frontiers | Frontiers in Sustainable Food Systems

TYPE Original Research
PUBLISHED 21 July 2025
DOI 10.3389/fsufs.2025.1612009

OPEN ACCESS

EDITED BY
Jianhui Ma,
Henan Normal University, China

REVIEWED BY
Rathinapriya Periyasamy,
National Institute of Horticultural and Herbal Science, Republic of Korea
Libei Li,
Zhejiang Agriculture and Forestry University, China

*CORRESPONDENCE
Jun Yan
xinsinian2006@163.com

RECEIVED 15 April 2025
ACCEPTED 07 July 2025
PUBLISHED 21 July 2025

CITATION
He J, Cui B, Liu P, Meng X and Yan J (2025) Utilizing machine learning and bioinformatics analysis to identify drought stress responsive genes in wheat (*Triticum aestivum* L.). *Front. Sustain. Food Syst.* 9:1612009. doi: 10.3389/fsufs.2025.1612009

Utilizing machine learning and bioinformatics analysis to identify drought stress responsive genes in wheat (*Triticum aestivum* L.)

Jiabei He[1,2], Baoyue Cui[1,2], Pingzeng Liu[1,2], Xianyong Meng[1,2] and Jun Yan[1,2]*

[1]College of Information Science and Engineering, Shandong Agricultural University, Tai'an, China, [2]Key Laboratory of Huang-Huai-Hai Smart Agricultural Technology, Ministry of Agriculture and Rural Affairs, Tai'an, China

One of the main abiotic stressors affecting agricultural output is drought stress, which has a substantial impact on wheat growth, development, and yield. This study aims to uncover transcriptomic changes in wheat leaves under drought stress using machine learning and bioinformatics approaches, thereby providing new research perspectives and solutions for understanding the mechanisms of abiotic stress responses in wheat and identifying drought-tolerant genes. First, publicly available RNA sequencing data on wheat drought stress were retrieved from databases, followed by sequence alignment and quantitative expression analysis. Differentially expressed genes (DEGs) under drought stress were identified through differential expression analysis. Subsequently, a weighted gene co-expression network was constructed to determine key gene modules, and multiple machine learning models were compared for their performance. Finally, an improved Random Forest-Boruta (RF-Boruta) algorithm was employed to identify key genes closely associated with drought stress responses. The differential expression analysis identified 16,754 DEGs, and the constructed gene co-expression network successfully identified modules related to drought stress responses. Among the various machine learning models, the random forest algorithm performed best in identifying drought stress-responsive genes. The improved RF-Boruta algorithm further selected candidate genes highly related to drought stress, improving model accuracy from 0.889 to 0.942 and the area under the curve (AUC) from 0.968 to 0.978. Gene enrichment analysis was also conducted. By integrating bioinformatics and machine learning techniques, this study identified key genes highly associated with drought stress responses in wheat, providing important insights into the potential mechanisms of drought responses in wheat.

KEYWORDS

drought stress, machine learning, transcriptome analysis, co-expression network, wheat

1 Introduction

Under natural agricultural conditions, crops are often affected by both biotic and abiotic stresses, with drought being one of the most severe abiotic stress factors (Seleiman et al., 2021). Over the centuries, drought has led to an average global yield reduction of cereal crops of approximately 13.7% (Lesk et al., 2016). As a result of land degradation and global warming, agricultural growth conditions are getting worse, and drought-related losses have overtaken those from all diseases combined (Gupta et al., 2020). According to earlier research, plants

under drought stress have lower rates of photosynthesis and CO_2 uptake, which has a negative impact on biomass buildup and production (Pinheiro and Chaves, 2011). In order to cope with the consequences of drought stress, plants have developed intricate molecular, physiological, and biochemical responses (Sun et al., 2020). Reducing water loss through stomata closure is a straightforward and efficient method of responding to drought stress. Numerous elements, such as osmotic regulation, environmental cues, photosynthesis, hormone regulation, CO_2 concentration, and respiration, influence how plants react to drought stress (Gong et al., 2020). Although a great deal of study has been done on the genes linked to drought tolerance, a thorough understanding of this complex feature in plants that is regulated by multiple genes still needs to be done (Kumar et al., 2017).

Given the complexity of drought stress responses and the need for a deeper understanding of the underlying genetic mechanisms, high-throughput technologies such as omics have become essential tools for identifying key genes and regulatory pathways involved in drought tolerance. Omics is a multidisciplinary field of study that uses a variety of high-throughput technologies to enable the qualitative and quantitative identification of genes, transcripts, transcription factors, and other parameters that have been previously identified or discovered (Ma et al., 2020). With advancements in high-throughput sequencing technology, research institutions worldwide are generating large-scale genomic data and depositing it in public domain databases (Yue et al., 2020). Understanding complicated biological systems in plants under biotic and abiotic stress is made possible by analyzing gene expression levels from these datasets across different trials. RNA sequencing (RNA-Seq) technology (Marguerat and Bähler, 2010) provides a comprehensive approach to studying gene functions and structures at a global level, revealing molecular mechanisms involved in specific biological processes. Additionally, RNA-Seq does not rely on predefined probe sets based on genome annotations, avoiding biases such as background noise introduced during microarray hybridization (Wang et al., 2009). Therefore, RNA-Seq technology is employed to perform comprehensive transcriptomic analyses.

With the development of high-quality reference genomes and the availability of large, publicly accessible RNA-Seq datasets (International Wheat Genome Sequencing Consortium et al., 2018), in-depth analyses of wheat RNA-Seq data to identify relevant genes have become feasible. In recent years, excellent chances for researching biological complexity have been made possible by omics data and artificial intelligence (such as machine learning and deep learning). Genes with comparable expression patterns that may be engaged in particular biological processes can be identified by studying gene co-expression networks, for instance, which offer an intuitive understanding of the interactions between chosen differentially expressed genes (DEGs) (Zhu et al., 2019). A potent technique for examining transcriptomic data and identifying patterns of gene co-expression is Weighted Gene Co-expression Network Analysis (WGCNA). Numerous plant species, including wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), Arabidopsis [*Arabidopsis thaliana* (L.) *Heynh.*], and maize (*Zea mays* L.), have had their transcriptional regulation successfully understood using this method (Chen et al., 2024; Yin et al., 2024; Guo et al., 2024; Liang et al., 2023).

Clarifying the characteristics of gene expression will influence our comprehension of plant physiology and contribute to increased crop yields (Jores et al., 2021). One of the several biological uses for supervised machine learning (ML) advancements is the prediction of gene expression patterns (Zrimec et al., 2020). In general, it falls into two categories: representation learning and classical machine learning. Even though deep learning has recently drawn a lot of attention to representation learning, classical machine learning still has some advantages, such as simpler model interpretation, easier configuration, lower computational requirements, and applicability to smaller datasets. The transcriptional regulation of abiotic stress in plants has been successfully studied in recent years using classical machine learning. To predict gene expression in rice under drought or high-temperature stress, for instance, a random forest machine learning model was trained (Smet et al., 2023). In order to evaluate the variable or consistent expression responses among maize genotypes, promoter characteristics and epigenetic markers were used to predict the genes in maize that respond to heat and cold stress (Zhou P. et al., 2022). Furthermore, by combining different approaches, machine learning methods can be utilized to find candidate genes. For instance, a machine learning method based on neural networks predicted stress-specific biomarker genes and stress types in plants (Kang et al., 2019). To find functional connections between genes and drought-specific transcription factors (TF), a web application using the machine learning technique Gene Regulation and Association Network (GRAiN) was recently developed. *OsbHLH148*, *OsRAP2.6*, *DREB1B*, *OsHSFA3*, *OsMYB6*, and *ONAC66* are among the drought-specific TFs that the online application effectively identified (Gupta et al., 2021). Furthermore, a number of plant-related applications have demonstrated the great potential of Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB) (Kumar et al., 2021).

In order to enable integrated studies, machine learning techniques use gene correlations to aggregate genes with comparable expression profiles under various experimental settings into co-expression modules (Zhu et al., 2024). These methods facilitate the identification of interconnected biological pathways and processes for DEGs. Machine learning is used to identify genes and narrow down the range of candidate genes. In this study, RNA-Seq gene expression data were used to analyze differential expression and co-expression networks to detect DEGs and key modules involved in drought stress responses. While the above approaches have proven effective for identifying key modules of DEGs (Hao et al., 2019), further identification through machine learning models is a critical step for selecting genes and predicting gene functions (Mahood et al., 2020). Therefore, we investigated the mechanisms underlying the identification of important genes linked to drought stress responses in wheat using machine learning techniques. These findings will lay a solid foundation for future research on drought tolerance-related genes in wheat.

2 Materials and methods

2.1 Wheat transcriptomic sequencing data collection

The overall design of this study was shown in Figure 1. The wheat raw sequencing datasets used in this study were obtained from the China National Center for Bioinformation (CNCB)[1] and National

1 https://www.cncb.ac.cn/

10.3389/fsufs.2025.1612009

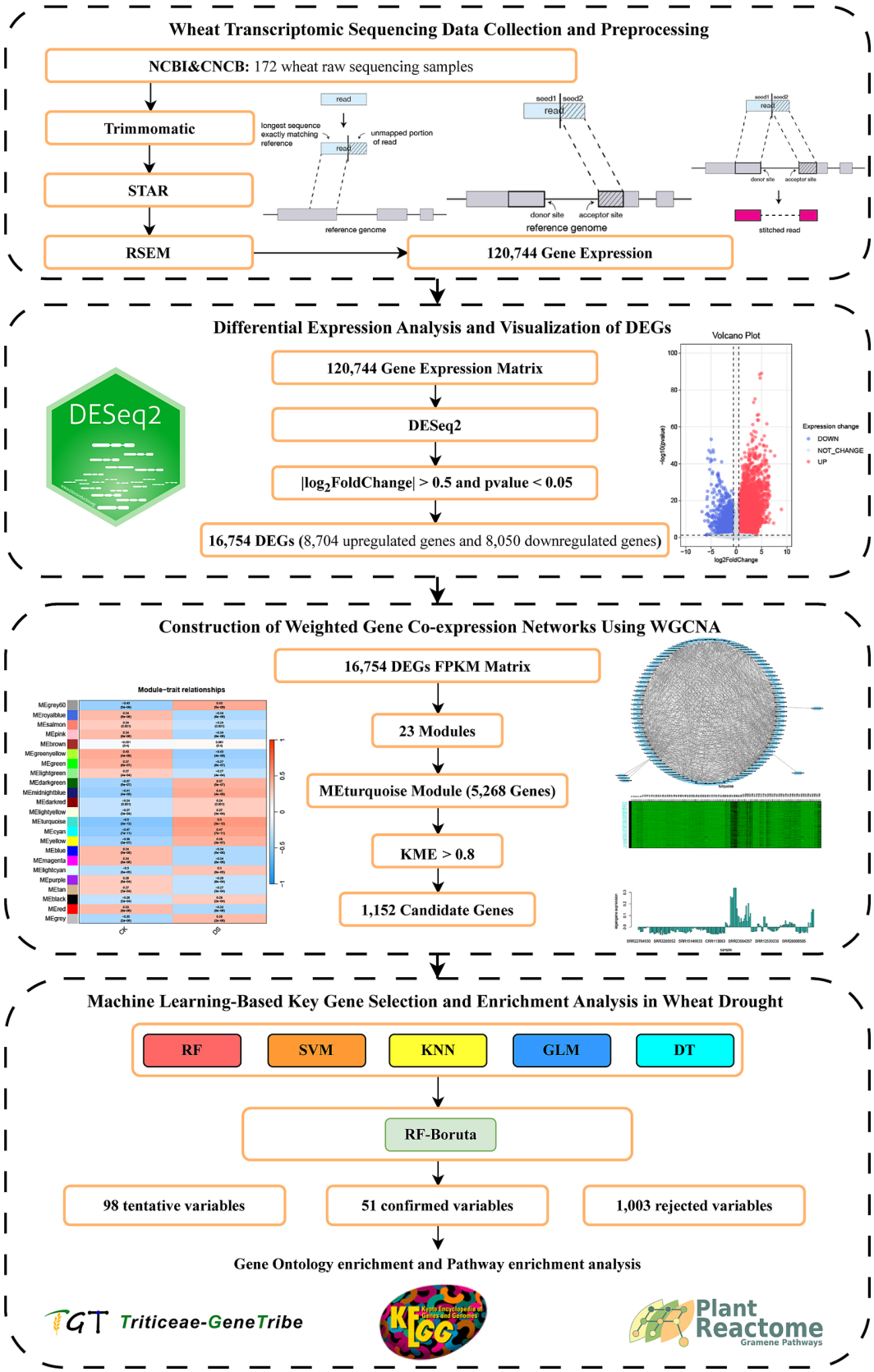

FIGURE 1
A flow-chart of this study.

Center for Biotechnology Information (NCBI)[2] databases, encompassing 172 wheat raw sequencing samples across 10 projects with project IDs: PRJCA002123, PRJCA013317, PRJNA316081, PRJNA659916, PRJNA746965, PRJNA838079, PRJNA911612, PRJNA914511, PRJNA937632, and PRJNA1077667. The treatment conditions and sampling tissues for PRJCA002123 were specified as follows: Wheat seedlings, aged 2 weeks, underwent water withdrawal, and their leaf transcriptomes were analyzed through RNA sequencing. The processing conditions and sampling organization for PRJCA013317 pertain to the transcriptome analysis of wheat seedling leaves subjected to drought stress conditions. The treatment conditions and sampling organizations for PRJNA316081 included WWC, which represented normal conditions [soil water potential (SWP) > −110 kPa], and drought conditions, characterized by a 24-h incubation following water withholding to achieve SWP < −390 kPa. This study utilized three biological replicates of RNA samples derived from the leaves of 40-day-old wheat. The study design for PRJNA659916 involved comparing gene expression variations in wheat flag leaves subjected to fully irrigated (wet) and water-deficient (dry) conditions. For PRJNA746965, the sampling organizations and treatment conditions were raw sequencing reads of winter wheat flag leaves under control and drought circumstances. PRJNA838079 was subjected to two treatment conditions and sampling organizations: drought and well-watered. The wheat leaf was employed for transcriptome analysis in both scenarios. For PRJNA911612, the sample organizations and treatment conditions were the transcriptomics of wheat leaves under CK and drought circumstances. For PRJNA914511, the treatment conditions and sampling organizations were CK-comparative transcriptome of leaves and wheat drought. PRJNA937632's treatment settings and sample arrangements included growing the wheat to GS13 (Zadok's growth scale), applying drought stress for 10 days (withholding water), and then recovering for 3 days (regular watering). The experiment was conducted using leaf tissue as the sampling tissue, with four replicates per treatment per landrace accession. Two wheat types' flag leaves were subjected to RNA sequencing as part of the PRJNA1077667 treatment conditions and sampling arrangements. Fastq files were downloaded from the CNCB database using the Edge turbo software, while SRR files from the NCBI database were downloaded and converted into fastq format using the SRA Toolkit (Leinonen et al., 2011). These datasets formed the basis for transcriptomic analysis. All 172 samples were divided into 86 control and 86 drought-treated groups, and all had at least three biological replications ($n = 3$) to ensure statistical robustness.

2.2 Quality control, read trimming, and transcriptome mapping

All downloaded data was subjected to quality inspection using the FastQC program. There are reads with poor sequencing quality and unusual sequencing lengths in the RNA-Seq raw data, such as reads shorter than 75. Moreover, the reads contain adaptor sequences that will subsequently impact the quantitative data of gene expression and must be filtered and eliminated beforehand using quality control. Sequences were trimmed and filtered using Trimmomatic software (Bolger et al., 2014) based on the following standards: SLIDINGWINDOW: 4:20, LEADING: 3, TRAILING: 3, MINLEN: 50. To obtain clean data, the results were converted to Phred-33. Two steps are needed to compare sequences using the Spliced Transcripts Alignment to a Reference (STAR) software (Dobin et al., 2013): Exons and shear sites were first extracted from the wheat reference genome and gene annotation files on the Ensembl website.[3] The reference genome and annotations were then used to create an index file based on the exon and shear site data that was obtained. It is feasible to identify the gene from which the transcription was carried out by using the index file as a control template for the base sequences in the ensuing sequence reads. In order to determine the greatest number of bases or amino acid residues that matched between the two genomes, the clean data were lastly compared to the wheat reference genome using STAR software. The STAR software comparison process uses the following filtering criteria: (1) The double-ended read length is filtered if one of the double-ended read lengths does not match; (2) The double-ended read length is filtered if its reference sequence differs from that of another double-ended read length; and (3) If there is a comparison between many positions, the double-ended read length will be filtered. Similarly, before calculating the expression, the RNA-Seq by Expectation Maximization (RSEM) (Li and Dewey, 2011) should build indexes based on the reference genome for wheat. To determine all of the gene expressions in each sample, RSEM software was used to quantify the Binary Alignment/Map format (BAM) data following STAR comparison.

2.3 Differential expression analysis and visualization of DEGs

The gene expression matrix (count values) obtained from RSEM was processed using a custom R script. Genes with expression levels of zero across all 172 samples, including both control and drought-treated groups, were filtered out. The remaining gene expression values were then rounded to integers. Differentially Expressed Genes (DEGs) were identified using the DESeq2 R package (Love et al., 2014), which normalizes the count data obtained from RSEM. The selection criteria for DEGs were set as $|\log_2\text{FoldChange}| > 0.5$ and p-value < 0.05. For the identified DEGs, principal component analysis (PCA) was performed based on their expression levels using the R packages FactoMineR (Lê et al., 2008) and factoextra. Volcano plots of DEGs were generated using DESeq2. The DEGs were then clustered to explore expression patterns. The optimal number of clusters was determined using the fviz_nbclust function from the cluster package. Finally, heatmaps of the DEGs were generated using the pheatmap package, illustrating the expression profiles across samples.

2 https://www.ncbi.nlm.nih.gov/

3 https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-59/fasta/triticum_aestivum/

2.4 Construction of weighted gene co-expression networks using WGCNA

The WGCNA R package (Langfelder and Horvath, 2008) was used to construct and analyze a weighted gene co-expression network based on sequencing data from 172 samples. Initially, the raw expression values (count data) of the differentially expressed genes were normalized to Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) values. Hierarchical clustering was then performed on all samples to identify and remove outlier samples. A soft threshold was selected by screening for an appropriate value, ensuring that the selected threshold approximates a scale-free network. The WGCNA package was then used to construct an undirected weighted gene co-expression network for the differentially expressed genes. The blockwiseModules function was employed to build the network with the following parameters: power = 7, minModuleSize = 50, and mergeCutHeight = 0.35. Correlation analyses were performed on the resulting co-expression modules, and a heatmap was generated to visualize these relationships. Core genes were identified from the network using the eigengene-based connectivity (KME) > 0.8, and a core gene dataset was constructed for further analysis.

2.5 Machine learning-based key gene selection and enrichment analysis in wheat drought response

To identify key genes responding to wheat drought stress, machine learning algorithms were applied using the gene expression matrix, core gene dataset, and sample data. The 172 samples were first divided into a training set (122 samples) and a test set (50 samples) in a 7:3 ratio using the caret package (Kuhn, 2008). The training set was further divided into a sub-training set (70%) and a test set (30%) for initial model development. Using the sub-training set, machine learning models, including Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Generalized Linear Model (GLM), and Decision Tree (DT), were trained with 5-fold cross-validation. These models were built using the randomForest, kernlab, and caret packages. Performance metrics for the models were compared and visualized using the DALEX package (Biecek, 2018). The best-performing model was selected for feature selection using the Boruta algorithm (Kursa et al., 2010). The Boruta algorithm, an ensemble method, identifies key features by iteratively testing the importance of features based on their contribution to classification accuracy. It does so by measuring the accuracy loss caused by random replacement of feature values. The refined feature subset was then used as the input matrix for the final training phase. Ten-fold cross-validation was used to retrain the model on the complete training set, and grid search was used to optimize hyperparameter tuning. The optimized model was applied to the independent test set for prediction. Finally, the model outputs were analyzed for Gene Ontology (GO) enrichment, gene annotation, and pathway enrichment using Triticeae-GeneTribe (TGT)[4] (Chen et al., 2020), Plant Reactome[5] (Naithani et al., 2020), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) online website[6] (Kanehisa and Goto, 2000), respectively. Setting parameters: p values were adjusted by the BH correction, and a false discovery rate (FDR) of 0.05 was applied to filter; the number of genes in the background ranges from 5 to 1,200. This analysis provided insights into the biological processes involving genes that respond to drought stress in wheat.

3 Results

3.1 Differential gene expression analysis of wheat RNA-seq under drought stress

Following quality control, low-quality repetitive sequences and splice sequences were clipped. To create the expression matrices, RMSE software was used to quantify 120,744 genes across all samples. From the initial 120,744 genes, 93,225 genes with non-zero expression levels across all samples were retained. Using the DESeq2 package, the raw gene expression counts were analyzed, resulting in the identification of 16,754 differentially expressed genes (DEGs) based on the set criteria. These DEGs included 8,704 upregulated genes and 8,050 downregulated genes (Supplementary Table 1). The distribution of differentially expressed genes at normalized expression levels in the control and drought-treated groups is displayed in the Minus-versus-Add (MA) plot prior to treatment. This plot represents the overall expression levels of the genes but has not yet been filtered for noise and adjusted for statistical significance (Figure 2A). Only significantly differentially expressed genes (red dots) were kept in the post-treatment MA plot after noise was removed and statistical significance was adjusted in comparison to the pre-treatment. Non-significant differences were eliminated, and the average expression level of genes was represented (Figure 2B).

The DEGs were subjected to principal component analysis (PCA) in order to determine whether the wheat samples showed unique expression profiles under various treatments (Figure 2C). In the PCA plot, CK represents the control group (wheat samples under normal growth conditions), while DS represents the drought treatment group (wheat samples subjected to drought stress). If the positions of the two samples in the PC1 and PC2 coordinate systems are very close to each other, it means that they have similar characteristics in gene expression profiles; on the contrary, if they are relatively far away from each other in the PC1 and PC2 coordinate systems, it means that their gene expression profiles are quite different. The CK and DS samples were clearly separated by the first principal component (PC1) and the second principal component (PC2), suggesting that the main cause of the significant variance in gene expression levels was the various treatments.

In order to illustrate the differential expression of genes according to the selection criteria, a volcano plot was finally created. As can be seen in the results of the volcano plot, more DEGs were upregulated than downregulated, with approximately more than 80% of the quantified genes not differentially expressed (Figure 2D). These results

4 http://wheat.cau.edu.cn/TGT/

5 https://plantreactome.gramene.org

6 http://www.kegg.jp/kegg/

10.3389/fsufs.2025.1612009

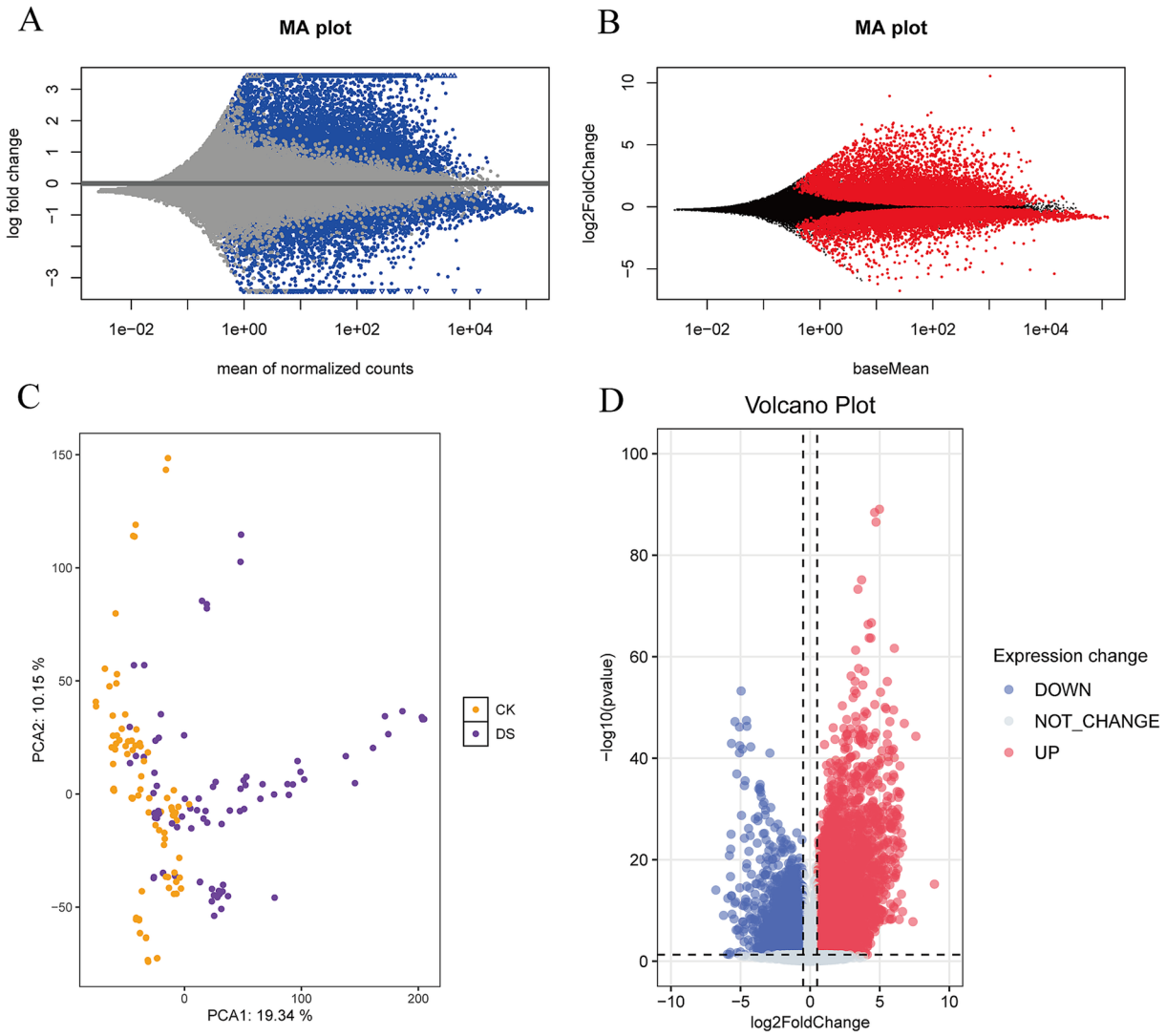

FIGURE 2
Transcriptome differential expression analysis. **(A)** MA plot with mean of normalized counts as horizontal coordinate and logarithm of multiplicity of differences as vertical coordinate. **(B)** MA plot with baseMean as the horizontal coordinate and the logarithm of the multiplicity of differences as the vertical coordinate. **(C)** Principal component analysis plot for CK and DS groups. CK represents the control group, while DS represents the drought treatment groups. **(D)** Volcano plot of differential genes. Where red represents up-regulated genes, blue represents down-regulated genes, and gray represents no change. *p*-value = 0.05 is indicated by the horizontal dashed line, and $\log_2$FoldChange = −0.5 and $\log_2$FoldChange = 0.5 are indicated by the vertical dashed lines.

demonstrate significant changes in gene expression between the control and drought-treated samples.

3.2 Clustering analysis of differentially expressed genes of wheat RNA-seq

For the identified 16,754 DEGs, the sum of squared error (SSE) within clusters was used to determine the optimal number of clusters. The goal was to minimize variability within clusters to group similar genes together while maximizing variability between clusters to separate dissimilar genes. The within-cluster sum of squares (WSS), which measures variation within a cluster, was plotted against the number of clusters. In the WSS plot, the number of clusters is shown on the x-axis, while WSS values are on the y-axis. A high WSS value indicates greater variability within clusters, while a low WSS value suggests well-defined clusters. As observed, the WSS dropped sharply between 1, 2, and 3 clusters, but the decrease became minimal after 3 clusters. Therefore, the optimal number of clusters was determined to be $k = 3$ (Supplementary Figure 1A).

A heatmap was then generated to visualize the expression patterns of DEGs within the three clusters. In the first cluster, genes exhibited an overall upregulation in the drought stress group compared to the control group. In contrast, genes in the second and third clusters showed an overall downregulation in the drought stress group (Supplementary Figure 1B). These trends in upregulation and downregulation were consistent with the general patterns observed in the volcano plot generated during the DESeq2 analysis. This clustering further highlights distinct gene expression patterns between control and drought-treated samples.

10.3389/fsufs.2025.1612009

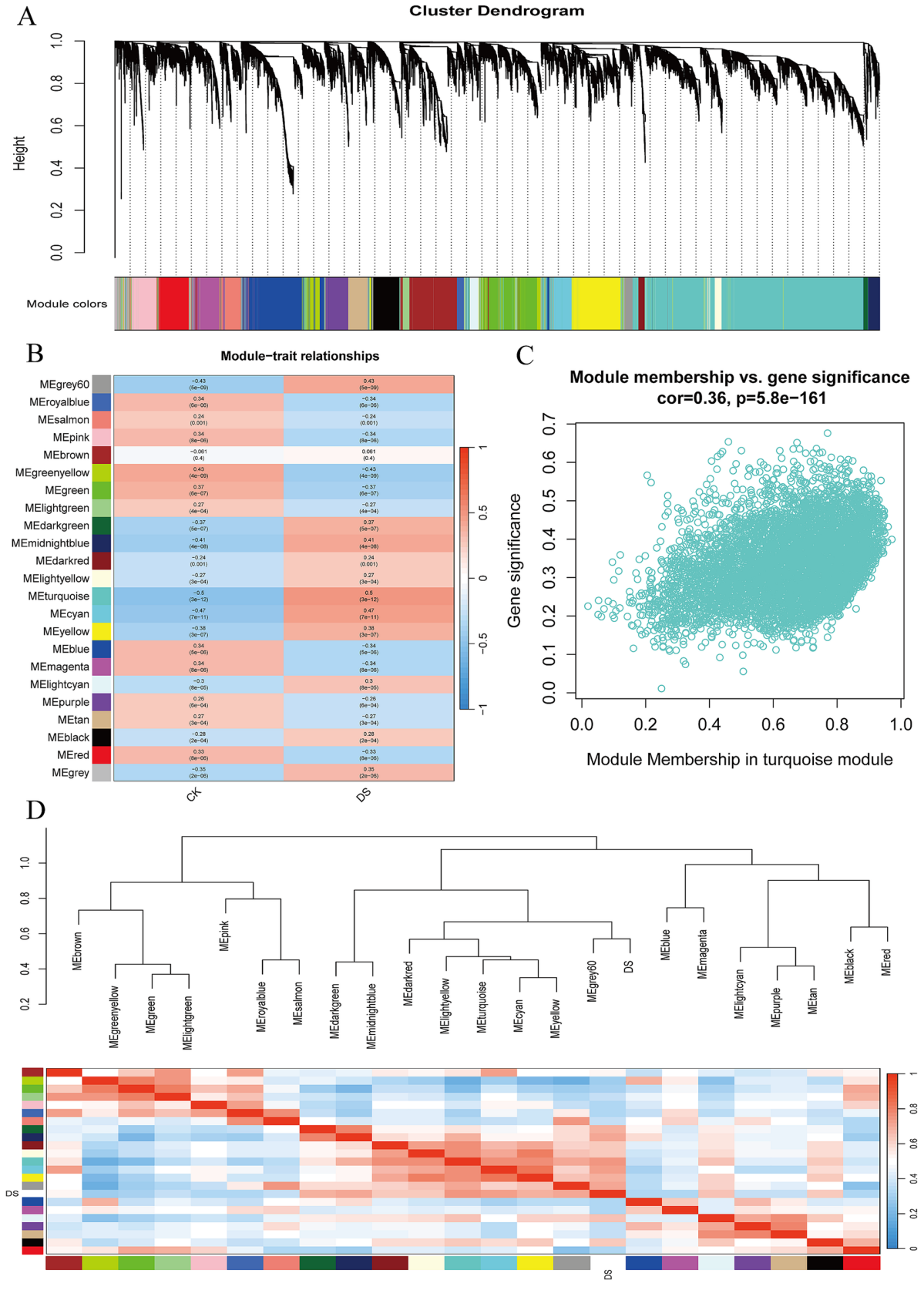

FIGURE 3
Drought stress-related modules identified. **(A)** Modular clustering tree. Each leaf node of the gene clustering tree represents a gene, and different clustering modules are distinguished from each other by colors. **(B)** Pearson correlation coefficients for modules and subgroups. Positive and negative

(Continued)

FIGURE 3 (Continued)
correlations are shown in red and blue, respectively, and the higher the correlation, the darker the color. **(C)** Scatterplot of MEturquoise modules. The horizontal coordinate is the correlation between genes and modules in the MEturquoise module, and the vertical coordinate is the correlation between genes and traits in the MEturquoise module. **(D)** Module correlation clustering tree and heat map. Each branch in the clustering tree represents a color module, and color modules with similar functionality are clustered into adjacent branches. The expression heatmap is a visualization of the correlation between modules; the higher the correlation, the redder the corresponding color in the heatmap, and vice versa, the bluer it is. The diagonal is the correlation between two modules that are the same, so it is 1.

3.3 Construction of a scale-free gene co-expression network via WGCNA

Weighted Gene Co-expression Network Analysis (WGCNA) was used to examine the traits of DEGs during drought stress. To create a co-expression module network linked to drought stress, all DEGs were incorporated. Initially, hierarchical clustering was used to identify outlier samples, ensuring a robust dataset for network analysis (Supplementary Figure 2A). WGCNA determines the soft threshold (β) by optimizing the scale-free topology of the network. A suitable β value improves the similarity of gene co-expression correlations, making the adjacency matrix calculation easier.

The weighted network graph meets the scale-free distribution, meaning that some nodes—referred to as hubs—have degrees that are noticeably higher than those of the general points, and some hubs are correlated with other nodes, which together make up the entire network. Following power function treatment, weak correlations take nth power, and the correlation drastically drops, while a few strong correlations are either unaffected or only slightly impacted in terms of gene expression associations. The soft threshold was determined using the scale-free topology fit index (R^2) and the mean connectivity of genes in the network for various β values. When the scale-free topology fit index R^2 was first more than 0.8, the value of $\beta = 7$ was used to build the weighted gene network (Supplementary Figures 2B,C; Supplementary Table 2).

The gene expression relationship can conform to a scale-free network, which has more low-degree nodes and fewer high-degree nodes, by selecting an appropriate value for β. A power law distribution governs the node degree k and the quantity of nodes with that degree. In particular, there is a negative correlation between the logarithm of the probability of the node's occurrence {log[p(k)]} and the logarithm of the number of nodes with degree k, log(k). To find out if the set parameter β satisfies the scale-free network, the frequency histograms of each degree of connection, k, and the distribution of log(k) vs. log[p(k)] are shown. The negative correlation between k and p(k), as illustrated in the figures, indicates that the selected β value ($\beta = 7$) can form a gene scale-free network; at this point, R^2 is 0.8 and the slope is −1.7 (Supplementary Figure 2D, E), offering a solid foundation for identifying the modules linked to drought stress.

3.4 Module detection and identification of drought-associated gene modules

Using the blockwiseModules function, an unsigned weighted co-expression network was constructed. Each leaf node of the gene clustering tree represents a gene, and different clustering modules are distinguished from each other by colors (Figure 3A; Supplementary Figure 3). The results categorized the DEGs into 23 distinct modules based on their expression similarity, with each module assigned a unique color. Genes in the grey module exhibited expression patterns dissimilar to all other modules, indicating no biological relevance, and were excluded from further analysis. Subsequent analyses focused on the remaining 22 modules and their associated genes.

The correlation between each module and the two experimental groups (CK for control and DS for drought treatment) was evaluated. A Pearson correlation heatmap revealed that the MEturquoise module showed the highest correlation with wheat drought stress ($p < 0.05$) (Figure 3B). A scatterplot of the MEturquoise module was generated to observe the distribution of genes within this module (Figure 3C; Supplementary Figure 4). Additionally, a dendrogram and heatmap of module correlations illustrated clustering relationships between the modules and the experimental conditions (Figure 3D). The MEturquoise module clustered with the drought stress treatment group (DS) in the same branch, reinforcing its association with drought stress.

The KME values for genes within the MEturquoise module were calculated (Supplementary Table 3). Based on the selection criteria, 1,152 genes from the MEturquoise module were identified for further analysis (Supplementary Figure 5). These genes are likely critical to understanding the molecular response to drought stress in wheat.

3.5 Evaluation of machine learning models for key genes

To identify key genes responding to drought stress in wheat, five machine learning models were constructed, and their performance was evaluated (Figure 4A). Residual analysis highlighted differences among the models. In the reverse cumulative distribution of absolute residuals (Figure 4B), it can be seen that the number of residuals is higher in the left tail of the KNN and GLM residual distributions, where KNN shows more large residuals than the other models, indicating that some predictions are less reliable. However, in the right tail of the residual distributions, GLM shows more large residual numbers than the other models, further indicating the difference in performance. Boxplots of residuals revealed that the DT model had the lowest median absolute residual, but the RF model performed best in terms of root mean square residual (RMSR) (Figure 4C). Histograms of residual distribution further demonstrated that the overall residuals of the RF model were significantly lower than those of the other four models (Figure 4D). Additionally, Receiver Operating Characteristic (ROC) curve analysis confirmed that the RF model achieved the highest performance (Figure 4E). Based on these comparisons, the RF model was selected, combined with the Boruta algorithm, for further analysis.

10.3389/fsufs.2025.1612009

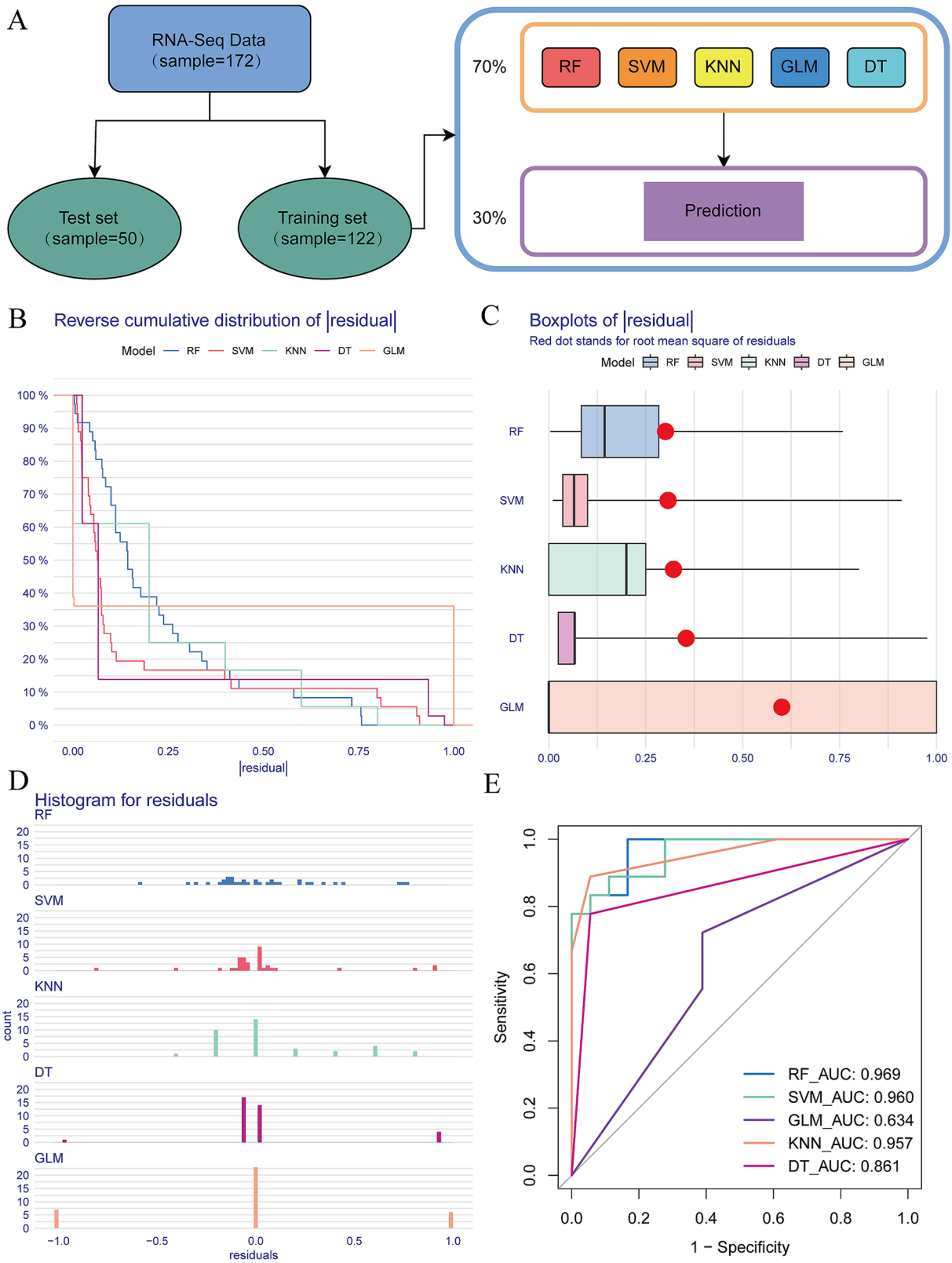

FIGURE 4
Machine learning model performance comparison. **(A)** Steps for dataset partitioning and comparative model selection. The dataset is 172 samples, which is divided into a training set of 122 samples and a test set of 50 samples according to the ratio of 7:3, where the 122-sample training set is again divided into a training set (70%) and a test set (30%). RF is the Random Forest model; SVM is the Support Vector Machine model; KNN is the K-adjacent model; GLM is the Generalized Linear Regression model; and DT is the Decision Tree model. **(B)** Reverse cumulative distribution of |residual|. The

(Continued)

FIGURE 4 (Continued)
horizontal coordinate is the absolute value of the residuals, and the vertical coordinate is the percentage number of samples. The closer the curve is to the lower left corner, the smaller the sample residuals are. **(C)** Boxplots of |residual|. The horizontal coordinate is the absolute value of the residuals, the red dot stands for the root mean square of the residuals, the box indicates that the data are discretely distributed, and the vertical line in the box is the median of the absolute value of the residuals. **(D)** Histogram for residuals. The horizontal coordinate is residuals, and the histogram shows the distribution of residuals for each model. **(E)** Model ROC curve and AUC. The ROC curve represents the model classification performance; the closer to the upper left performance, the better the classification performance. The AUC value is the area under the ROC curve to quantify the ROC curve.

3.6 Feature selection and functional analysis of key drought-responsive genes

After 100 iterations, the importance scores of variables stabilized. Confirmation variables (green) have small changes in importance ranking and scores after 50 iterations, indicating that the model has been basically stable. Tentative variables (yellow) have fluctuations, but the overall distribution is more consistent. Fluctuations in rejected variables (red) do not affect the screening of key genes. Shadow variables (blue) fluctuate less, indicating that the random noise has limited influence. The Boruta algorithm identified 51 confirmed, 98 tentative, and 1,003 rejected variables. A plot of variable importance over iterations showed green for confirmed variables, yellow for tentative variables, red for rejected variables, and blue indicating the minimum, mean, and maximum importance of shadow variables (Figure 5A; Supplementary Table 4).

The RF model was then optimized by integrating the Boruta algorithm. Fixed hyperparameters were used in the RF model, and Boruta was applied to select features from the training gene expression matrix. This selected feature subset was used to train a new RF model, which was further optimized through grid search. Comparing the performance of the Random Forest-Boruta (RF-Boruta) model with the original RF model, the RF-Boruta model outperformed in all evaluated aspects (Figures 5B–D; Table 1).

Finally, the variable importance scores identified by Boruta were visualized (Figure 5E). Gene Ontology (GO) enrichment analysis, gene annotation, and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis were conducted for the confirmed variables, revealing their biological significance and their roles in wheat's response to drought stress.

The genes GO enrichment analysis of the confirmed genes revealed significant associations with biological processes (BP) and molecular functions (MF) relevant to drought stress and plant development. In the BP category, the confirmed genes were enriched in terms such as carotene catabolic process (GO:0015121), abscisic acid biosynthetic process (GO:0009688), and abscisic acid-activated signaling pathway (GO:0009738), highlighting their roles in ABA synthesis and signaling under stress. Other enriched terms included post-embryonic plant organ morphogenesis (GO:0090697), linked to organ development under stress, and negative regulation of peroxidase activity (GO:2000469), indicating modulation of reactive oxygen species (ROS). In the MF category, the confirmed genes were involved in critical enzymatic activities and binding functions such as magnesium-dependent protein serine/threonine phosphatase activity (GO:0004724), essential for signal transduction; 9-cis-epoxycarotenoid dioxygenase activity (GO:0045549), key in ABA biosynthesis; as well as disaccharide binding (GO:0048030) and phosphatidylcholine binding (GO:0031210), associated with osmotic regulation and membrane signaling (Figure 6; Table 2; Supplementary Table 5). These findings underscore the importance of the identified genes in mediating wheat's response to drought stress, providing valuable targets for improving drought tolerance.

Pathway enrichment analysis showed that the confirmed genes were mainly involved in the MAPK signaling pathway-plant, biosynthesis of secondary metabolites, plant hormone signal transduction, carotenoid biosynthesis, abscisic acid biosynthesis and mediated signaling, abscisic acid homeostasis, and so on (Supplementary Table 6). Protein phosphatase 2C (*PP2C*) is involved in the MAPK signaling pathway—plant adaptation under the influence of abscisic acid from salt/drought/osmotic stress in plants—and it prevents downstream signaling by dephosphorylating Sucrose Non-Fermenting 1-Related Protein Kinase 2 (*SnRK2*) kinase, whereas the presence of abscisic acid (*ABA*) inhibits PP2C, which allows the activation of the stress signaling pathway (Supplementary Figure 6). Autoinhibited H + -ATPase (*AHA1/2*) activation is regulated by growth hormone (auxin). Auxin is transported into the cell via AUX1 and activates downstream signaling pathways (e.g., *ABP1* and *TMK1/4*) that ultimately activate AHA1/2. ABA Insensitive (*ABI1/2*) acts as a negative regulator, inhibiting the activity of SnRK2 kinase and thus blocking ABA signaling. ABA-responsive element-binding factor (*ABF*) acts as a transcription factor, directly regulating the expression of downstream genes. In signaling, ABA activates Pyrabactin Resistance/Pyrabactin Resistance-Like (*PYR/PYL*); PYR/PYL inhibits PP2C and deregulates the inhibition of SnRK2, which is then activated and phosphorylated by downstream targets (Supplementary Figure 7). 2.5.1.32 (Phytoene Synthase, *PSY*) catalyzes the condensation reaction of geranylgeranyl diphosphate (*GGPP*) molecules to produce prephytoene diphosphate (*PPP*), which is subsequently isomerized to phytoene; 1.13.11.51 (9-cis-epoxycarotenoid dioxygenase, *NCED*) generates xanthoxin and other by-products by catalyzing the cleavage of 9-cis-epoxycarotenoids (e.g., 9-cis-violaxanthin or 9-cis-neoxanthin) for further conversion to abscisic aldehyde, and finally ABA is synthesized in several steps (Supplementary Figure 8).

4 Discussion

One of the biggest environmental stresses is drought, which restricts plant growth, development, and reproduction and poses a major risk to food security and agricultural productivity worldwide (Kumar et al., 2023; Shahriari et al., 2022; Waititu et al., 2021). Several drought-tolerant genes in wheat, such as *TaNAC071-A, TaSAP5,* and *TaSNAC8-6A,* have been identified in previous studies (Mao et al., 2022; Zhang et al., 2017; Mao et al., 2020). In this study, we looked at possible characterization genes linked to drought stress response in

10.3389/fsufs.2025.1612009

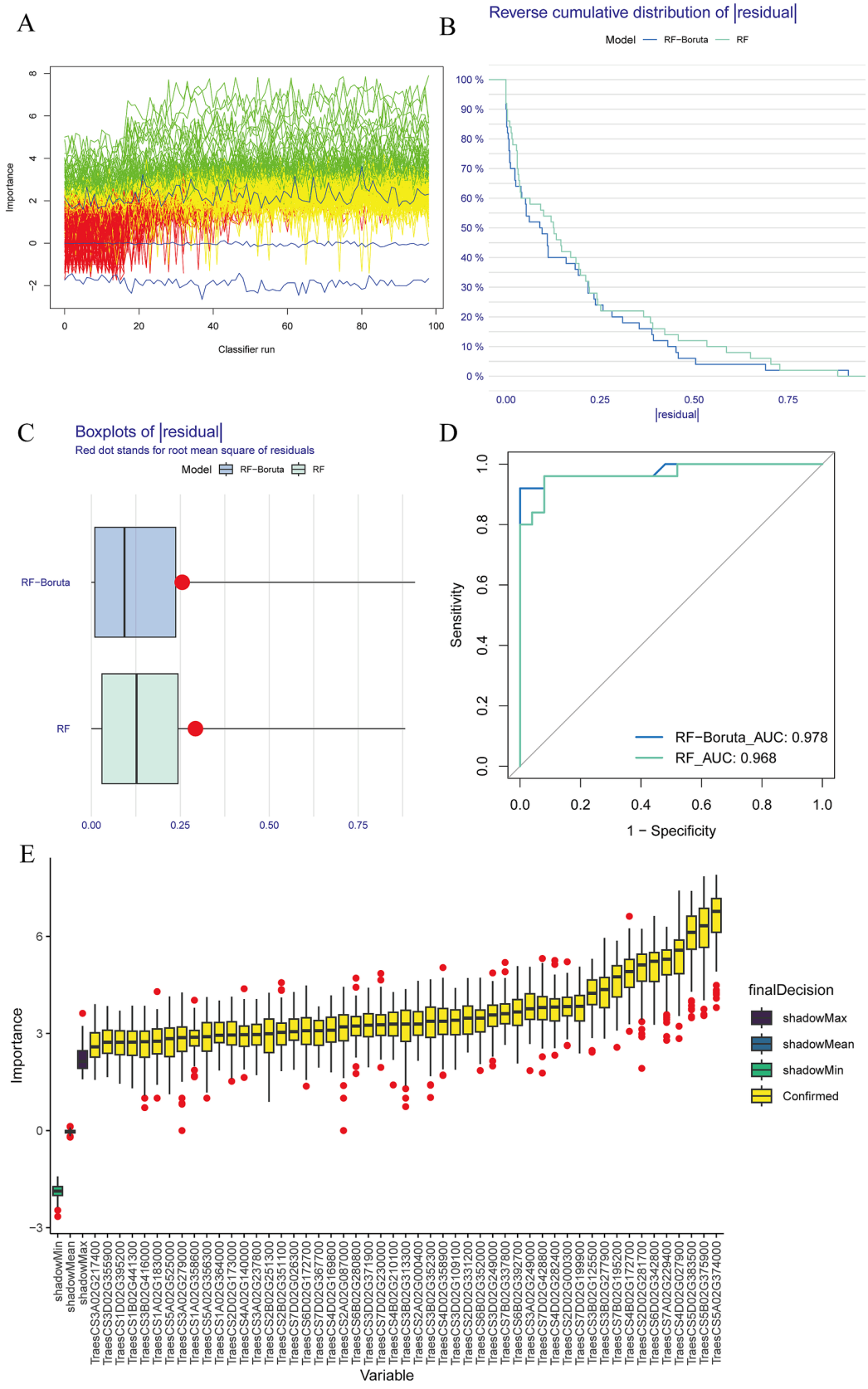

FIGURE 5
The RF-Boruta algorithm identifies key genes for drought stress. **(A)** Boruta algorithm variable importance scores with 100 iterations. Confirmed variables in green, tentative variables in yellow, rejected variables in red, and the importance of the minimum, average, and maximum shadow variables in blue, respectively. **(B)** Reverse cumulative distribution of |residual|. **(C)** Boxplots of |residual|. **(D)** Model ROC curve and AUC. **(E)** Order of importance of confirmed variables. Confirmed variables are in yellow; the rest are minimum, average, and maximum shadow variables.

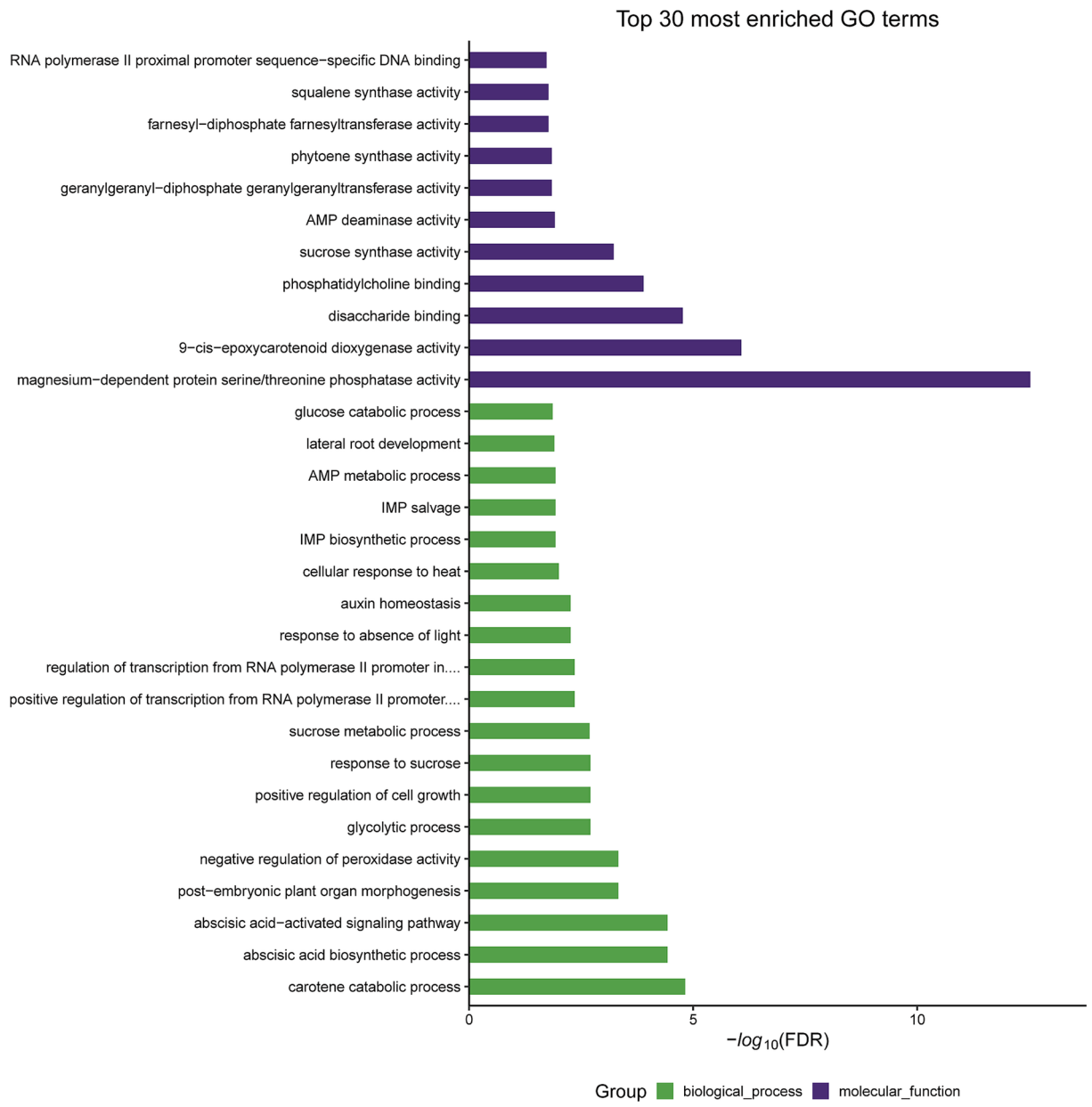

FIGURE 6
Confirmed variant gene ontology enrichment analysis. The bars show the results of the top 30 sorted GO terms, with the green bar indicating the biological process, the purple bar indicating the molecular function, and the text on the left indicating the description corresponding to each GO term.

TABLE 1 Model comparison parameters, results retained to three decimal places.

Model	Accuracy	F1	Recall	AUC
RF	0.889	0.864	0.766	0.968
RF-Boruta	0.942	0.936	0.886	0.978

wheat by combining bioinformatics approaches with improved machine learning techniques.

Our feature selection results based on core DEGs using machine learning algorithms revealed associations with multiple biological processes, including the carotene catabolic process, abscisic acid biosynthetic process, abscisic acid-activated signaling pathway, post-embryonic plant organ morphogenesis, and negative regulation of peroxidase activity. Furthermore, molecular function analysis uncovered roles in magnesium-dependent protein serine/threonine phosphatase activity, 9-cis-epoxycarotenoid dioxygenase activity, disaccharide binding, and the mitochondrial pyruvate dehydrogenase complex. Some of our gene annotation highlighted genes such as *MYB* and *bZIP* transcription factors (*TraesCS3B02G125500, TraesCS3D02G109100, TraesCS3D02G371900*), which are modulated during plant drought responses; they have also been shown to be drought-resistant genes in other plants (including Arabidopsis, rice, and soybean) (Mao et al., 2020). *NCED*, 9-cis-epoxycarotenoid dioxygenase, and protein phosphatase 2C (*TraesCS5A02G374000,*

10.3389/fsufs.2025.1612009

TABLE 2 Confirmed variant gene annotation.

Gene	Description	*Arabidopsis thaliana*
TraesCS5A02G374000	9-cis-epoxycarotenoid dioxygenase NCED3, chloroplastic	*AT3G14440*
TraesCS5B02G375900	9-cis-epoxycarotenoid dioxygenase NCED3, chloroplastic	*AT3G14440*
TraesCS5D02G383500	9-cis-epoxycarotenoid dioxygenase NCED3, chloroplastic	*AT1G30100*
TraesCS4D02G027900	Plasma membrane ATPase 1	*AT5G62670*
TraesCS7A02G229400	Protein MOTHER of FT and TFL1 homolog 1	*AT1G18100*
TraesCS6D02G342800	Protein LAZ1 homolog 2	*AT1G23070*
TraesCS4B02G172700	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	*AT3G08590*
TraesCS7B02G195200	Protein MOTHER of FT and TFL1 homolog 1	*AT1G18100*
TraesCS3B02G277900	Protein phosphatase 2C 50	*AT4G26080*
TraesCS3B02G125500	Transcription factor MYBS3	*AT5G56840*
TraesCS7D02G199900	Pyruvate dehydrogenase E1 component subunit alpha-2, mitochondrial	*AT1G59900*
TraesCS2D02G000300	Probable protein phosphatase 2C 37	NA
TraesCS4D02G282400	Embryonic protein DC-8	NA
TraesCS7D02G428800	18 kDa seed maturation protein	NA
TraesCS3A02G249000	Protein phosphatase 2C 50	*AT1G72770*
TraesCS6B02G392700	Protein LAZ1 homolog 2	*AT1G23070*
TraesCS7B02G337800	18 kDa seed maturation protein	NA
TraesCS3D02G249000	Protein phosphatase 2C 50	NA
TraesCS2D02G331200	ATP-dependent 6-phosphofructokinase 2	*AT5G47810*
TraesCS3D02G109100	Transcription factor MYBS3	*AT5G56840*
TraesCS4D02G358900	Heat stress transcription factor C-2a	NA
TraesCS3B02G352300	Receptor-like serine/threonine-protein kinase SD1-8	*AT1G61610*
TraesCS2A02G000400	Probable protein phosphatase 2C 37	NA
TraesCS4B02G210100	Probable protein phosphatase 2C 30	*AT2G29380*
TraesCS7D02G230000	Protein MOTHER of FT and TFL1 homolog 1	*AT1G18100*
TraesCS3D02G371900	bZIP transcription factor 12	NA
TraesCS6B02G280800	Ethylene-responsive transcription factor 5	*AT5G47230*
TraesCS2A02G087000	Probable AMP deaminase	*AT2G38280*
TraesCS4D02G169800	Sucrose synthase 4	*AT4G02280*
TraesCS7D02G367700	Heat stress transcription factor C-2b	*AT2G26150*
TraesCS6D02G172700	CASP-like protein 2D1	NA
TraesCS7D02G026300	Late embryogenesis abundant protein 6	NA
TraesCS2B02G351100	Amino acid permease 2	*AT5G23810*
TraesCS3A02G237800	Probable protein phosphatase 2C 8	*AT4G11040*
TraesCS4A02G140000	Sucrose synthase 4	*AT4G02280*
TraesCS2D02G173000	Fasciclin-like arabinogalactan protein 2	NA
TraesCS1A02G364000	Late embryogenesis abundant protein, group 3	NA
TraesCS5A02G356300	Phytoene synthase 3, chloroplastic	*AT5G17230*
TraesCS1A02G358600	Protein phosphatase 2C 50	*AT5G57050*
TraesCS5A02G525000	F-box protein At4g00755	NA
TraesCS1A02G183000	Probable trehalase	NA
TraesCS3B02G416000	Probable magnesium transporter NIPA6	*AT4G38730*
TraesCS1B02G441300	Ethylene-responsive transcription factor ERF061	NA
TraesCS1D02G395200	Aspartic proteinase NANA, chloroplast	NA
TraesCS3D02G355900	Probable protein phosphatase 2C 9	*AT2G29380*

The last column is a homologous gene of *Arabidopsis thaliana*. **TraesCS2D02G281700*, *TraesCS6B02G352000*, *TraesCS3B02G313300*, *TraesCS2B02G251300*, *TraesCS3A02G279000*, and *TraesCS3A02G217400* No longer shown due to missing annotations.

TraesCS5B02G375900, TraesCS5D02G383500, TraesCS3B02G277900, TraesCS3A02G249000, TraesCS3D02G249000, TraesCS4B02G210100, TraesCS1A02G358600, TraesCS3D02G355900) are involved in drought stress response as a phytohormone pathway, demonstrating that under severe drought conditions, enhanced drought adaptability depends on ABA (Niu et al., 2023; Shi et al., 2015); Plasma membrane ATPase (*TraesCS4D02G027900*) is involved in the regulation of abscisic acid (ABA)-induced stomatal closure under drought stress. Regulation of plasma membrane H-ATPase activity by identification of *ZmMHA2*, a homolog of Arabidopsis PM H-ATPase protein, and *ZmCRK1*, a CDPK-related kinase; reduced *ZmCRK*1 abundance during drought conditions lessens its inhibitory action on *ZmMHA2*, increasing H efflux and causing stomatal closure in response to drought stress (Liu et al., 2024); Receptor-like serine/threonine-protein kinase (*TraesCS3B02G352300*) regulates the regulation of plant growth and development during abiotic stress (Gandhi and Oelmüller, 2023); 2,3-bisphosphoglycerate-independent phosphoglycerate mutase (*TraesCS4B02G172700*) is involved in the glycolytic pathway in soybean roots under drought stress. It was demonstrated that improved carbohydrate metabolism leads to more energy being produced in the cell, which in turn increases the soybean root system's ability to withstand drought under osmotic stress (Zhou Y. et al., 2022); ATP-dependent 6-phosphofructokinase 2 (*TraesCS2D02G331200*) appears in a pathway model of wheat seedlings in response to drought stress (Wang et al., 2019); Aspartic proteinase (*TraesCS1D02G395200*) expression can directly or indirectly promote the expression of ABA-related genes. Get through interacting with different drought-associated proteins, especially those related to stomatal closure and density, to confer drought tolerance in *Arabidopsis thaliana* (Fernando et al., 2020). For hormone homeostasis and drought stress tolerance, the amino acid permease (*TraesCS2B02G351100*) gene is essential. Eight phytohormone-responsive factors and four drought stress response factors are found in maize *ZmAAP1*. Increased expression of some *ZmAAPs* and *ZmCATs* in leaves under drought stress indicates the significance they play in drought stress acclimation and resistance (Islam et al., 2024). The functional significance of the aforementioned genes was further highlighted using the machine learning-based feature selection technique.

The ability to integrate and analyze large volumes of data to extract accurate information is crucial for precision breeding in crops and increasing agricultural yields in unfavorable settings, as evidenced by the rapid improvement of sequencing technology and the growing demand for multi-omics integrated analysis. Because of its remarkable capacity and adaptability in combining many types of biological knowledge and omics data, machine learning, a branch of artificial intelligence, exhibits significant promise in this sector. In this study, we used machine learning, which offered a strong framework for deciphering intricate gene expression data. In order to discover core DEGs and their functions in drought stress response, we compared a number of machine learning approaches, such as Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Generalized Linear Model (GLM), and Decision Tree (DT) models. The RF model, known for its accuracy and efficiency, was combined with the Boruta algorithm to enhance feature selection. By introducing shadow variables for iterative comparisons, Boruta identified 51 significant genes with clear contributions to drought response pathways. The massive dataset of 16,754 DEGs linked to drought treatments that was gathered using conventional approaches is simplified by this method, which also makes data processing easier. This integration improved model accuracy from 0.889 to 0.942 and area under the curve (AUC) from 0.968 to 0.978, highlighting these genes' importance in drought tolerance. These identified genes, derived from integrating multiple wheat varieties, hold potential for application in breeding programs to improve drought resilience.

In addition to identifying important genes involved in the drought stress response, this study showed how bioinformatics analysis and machine learning techniques may be applied in agricultural genomics. The genes that were found, particularly those that the RF-Boruta model highlighted, could be used as biomarkers to screen for and create drought-resistant wheat cultivars. Notably, the model's ability to generalize across datasets is critical to its utility in real-world breeding programs. Therefore, ensuring that training data are diverse and representative of various wheat genotypes, growth stages, and environmental conditions is essential for building robust predictive models. Future studies should focus not only on experimentally validating these candidate genes but also on integrating additional sequencing datasets from different crop species and stress scenarios to enhance model robustness. Expanding the dataset spectrum will improve the model's generalization capacity, reduce overfitting, and facilitate the discovery of stress-resilient genes across species. Ultimately, by combining high-throughput genomic data with advanced computational methods, crop varieties adaptable to a wide range of environmental conditions can be developed more efficiently—contributing to both food security and agricultural sustainability.

5 Conclusion

This study establishes an effective and integrative framework for identifying drought-responsive genes in wheat by combining transcriptomic analysis with machine learning techniques. By systematically comparing multiple models and optimizing the Random Forest algorithm with Boruta feature selection, we achieved two key improvements:

(1) Enhanced gene prioritization–The RF-Boruta method effectively reduced the dimensionality of the dataset from 16,754 to 51 core genes with strong predictive power, thereby simplifying downstream analyses and improving interpretability.

(2) Improved predictive performance–Through refined feature selection and model optimization, the prediction accuracy improved from 0.889 to 0.942, and the AUC increased from 0.968 to 0.978, underscoring the robustness and precision of our framework in identifying wheat-relevant drought-responsive genes.

These advances not only improve the efficiency and interpretability of gene discovery but also provide valuable candidates for breeding drought-resilient wheat varieties. The framework offers broad applicability for multi-omics integration and stress-response studies in crop genomics. In the end, the characterized genes help to increase the sustainability and productivity of wheat production in unfavorable environmental conditions by serving as valuable references for breeding programs with enhanced drought resistance.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JH: Conceptualization, Writing – original draft, Methodology, Visualization. BC: Writing – review & editing, Investigation, Data curation. PL: Writing – review & editing, Investigation. XM: Investigation, Writing – review & editing. JY: Methodology, Supervision, Conceptualization, Writing – review & editing, Validation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by funding from Jun Yan: A Project of Shandong Province Higher Educational Program for Introduction and Cultivation of Young Innovative Talents in 2021.

Acknowledgments

We thank all contributors for their work and would like to thank the editor and reviewers for their valuable comments and suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsufs.2025.1612009/full#supplementary-material

References

Biecek, P. (2018). DALEX: explainers for complex predictive models in R. *J. Mach. Learn. Res.* 19, 1–5. doi: 10.48550/arXiv.1806.08915

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chen, Y., Song, W., Xie, X., Wang, Z., Guan, P., Peng, H., et al. (2020). A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the triticeae tribe as a pilot practice in the plant pangenomic era. *Mol. Plant* 13, 1694–1708. doi: 10.1016/j.molp.2020.09.019

Chen, J., Zhang, L., Liu, Y., Shen, X., Guo, Y., Ma, X., et al. (2024). RNA-Seq-based WGCNA and association analysis reveal the key regulatory module and genes responding to salt stress in wheat roots. *Plan. Theory* 13:274. doi: 10.3390/plants13020274

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Fernando, F. D., Raúl, D. G., and Gabriela, G. M. (2020). Overexpression of Arabidopsis aspartic protease APA1 gene confers drought tolerance. *Plant Sci.* 292:110406. doi: 10.1016/j.plantsci.2020.110406

Gandhi, A., and Oelmüller, R. (2023). Emerging roles of receptor-like protein kinases in plant response to abiotic stresses. *Int. J. Mol. Sci.* 24:14762. doi: 10.3390/ijms241914762

Gong, L., Zhang, H., Liu, X., Gan, X., Nie, F., Yang, W., et al. (2020). Ectopic expression of HaNAC1, an ATAF transcription factor from Haloxylon ammodendron, improves growth and drought tolerance in transgenic Arabidopsis. *Plant Physiol. Biochem.* 151, 535–544. doi: 10.1016/j.plaphy.2020.04.008

Guo, Y., Liu, C., Zhang, Y., Zheng, S., Cao, P., Wang, X., et al. (2024). Characterization key genes of Arabidopsis seedlings in response to β-caryophyllene, eugenol using combined transcriptome and WGCN analysis. *Front. Plant Sci.* 14:1295779. doi: 10.3389/fpls.2023.1295779

Gupta, A., Rico-Medina, A., and Caño-Delgado, A. I. (2020). The physiology of plant responses to drought. *Science* 368, 266–269. doi: 10.1126/science.aaz7614

Gupta, C., Ramegowda, V., Basu, S., and Pereira, A. (2021). Using network-based machine learning to predict transcription factors involved in drought resistance. *Front. Genet.* 12:652189. doi: 10.3389/fgene.2021.652189

Hao, Z., Geng, M., Hao, Y., Zhang, Y., Zhang, L., Wen, S., et al. (2019). Screening for differential expression of genes for resistance to *Sitodiplosis mosellana* in bread wheat via BSR-seq analysis. *Theor. Appl. Genet.* 132, 3201–3221. doi: 10.1007/s00122-019-03419-9

International Wheat Genome Sequencing ConsortiumAppels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Islam, M. N., Rabby, M. G., Hossen, M. M., Bonny, M., and Hasan, M. M. (2024). Genome-wide identification following functional analysis of amino acid permease and cationic amino acid transporter gene families in maize and their role in drought stress. *S. Afr. J. Bot.* 168, 360–371. doi: 10.1016/j.sajb.2024.03.029

Jores, T., Tonnies, J., Wrightsman, T., Buckler, E. S., Cuperus, J. T., Fields, S., et al. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* 7, 842–855. doi: 10.1038/s41477-021-00932-y

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kang, D., Ahn, H., Lee, S., Lee, C. J., Hur, J., Jung, W., et al. (2019). StressGenePred: a twin prediction model architecture for classifying the stress types of samples and discovering stress-related genes in arabidopsis. *BMC Genomics* 20:949. doi: 10.1186/s12864-019-6283-z

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Kumar, N., Mishra, B. K., Liu, J., Mohan, B., Thingujam, D., Pajerowska-Mukhtar, K. M., et al. (2023). Network biology analyses and dynamic modeling of gene regulatory networks under drought stress reveal major transcriptional regulators in arabidopsis. *Int. J. Mol. Sci.* 24:7349. doi: 10.3390/ijms24087349

Kumar, R., Baloch, G., Buriro, A. B., and Bhatti, J. (2021). Fungal blast disease detection in rice seed using machine learning. *Int. J. Adv. Comput. Sci. Appl.* 12, 248–258. doi: 10.14569/IJACSA.2021.0120232

Kumar, S., Beena, A., Awana, M., and Singh, A. (2017). Physiological, biochemical, epigenetic and molecular analyses of wheat (*Triticum aestivum*) genotypes with contrasting salt tolerance. *Front. Plant Sci.* 8:1151. doi: 10.3389/fpls.2017.01151

Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundam. Inf.* 101, 271–285. doi: 10.3233/FI-2010-288

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Lê, S., Josse, J., and Husson, F. (2008). Factominer: an R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/jss.v025.i01

Leinonen, R., Sugawara, H., and Shumway, M.International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Lesk, C., Rowhani, P., and Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature* 529, 84–87. doi: 10.1038/nature16467

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Liang, T., Hu, Y., Xi, N., Zhang, M., Zou, C., Ge, F., et al. (2023). GWAS across multiple environments and WGCNA suggest the involvement of ZmARF23 in embryonic callus induction from immature maize embryos. *Theor. Appl. Genet.* 136:93. doi: 10.1007/s00122-023-04341-x

Liu, J., Li, X. D., Jia, D., Qi, L., Jing, R., Hao, J., et al. (2024). ZmCRK1 negatively regulates maize's response to drought stress by phosphorylating plasma membrane H+-ATPase ZmMHA2. *New Phytol.* 244, 1362–1376. doi: 10.1111/nph.20093

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Ma, X., Meng, Y., Wang, P., Tang, Z., Wang, H., and Xie, T. (2020). Bioinformatics-assisted, integrated omics studies on medicinal plants. *Brief. Bioinform.* 21, 1857–1874. doi: 10.1093/bib/bbz132

Mahood, E. H., Kruse, L. H., and Moghe, G. D. (2020). Machine learning: a powerful tool for gene function prediction in plants. *Appl. Plant Sci.* 8:e11376. doi: 10.1002/aps3.11376

Mao, H., Li, S., Chen, B., Jian, C., Mei, F., Zhang, Y., et al. (2022). Variation in cis-regulation of a NAC transcription factor contributes to drought tolerance in wheat. *Mol. Plant* 15, 276–292. doi: 10.1016/j.molp.2021.11.007

Mao, H., Li, S., Wang, Z., Cheng, X., Li, F., Mei, F., et al. (2020). Regulatory changes in TaSNAC8-6A are associated with drought tolerance in wheat seedlings. *Plant Biotechnol. J.* 18, 1078–1092. doi: 10.1111/pbi.13277

Marguerat, S., and Bähler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* 67, 569–579. doi: 10.1007/s00018-009-0180-6

Naithani, S., Gupta, P., Preece, J., D'Eustachio, P., Elser, J. L., Garg, P., et al. (2020). Plant Reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.* 48, D1093–D1103. doi: 10.1093/nar/gkz996

Niu, Y., Li, J., Sun, F., Song, T., Han, B., Liu, Z., et al. (2023). Comparative transcriptome analysis reveals the key genes and pathways involved in drought stress response of two wheat (*Triticum aestivum* L) varieties. *Genomics* 115:110688. doi: 10.1016/j.ygeno.2023.110688

Pinheiro, C., and Chaves, M. M. (2011). Photosynthesis and drought: can we make metabolic connections from available data? *J. Exp. Bot.* 62, 869–882. doi: 10.1093/jxb/erq340

Seleiman, M. F., Al-Suhaibani, N., Ali, N., Akmal, M., Alotaibi, M., Refay, Y., et al. (2021). Drought stress impacts on plants and different approaches to alleviate its adverse effects. *Plan. Theory* 10:259. doi: 10.3390/plants10020259

Shahriari, A. G., Soltani, Z., Tahmasebi, A., and Poczai, P. (2022). Integrative system biology analysis of transcriptomic responses to drought stress in soybean (*Glycine max* L.). *Genes.* 13:1732. doi: 10.3390/genes13101732

Shi, L., Guo, M., Ye, N., Liu, Y., Liu, R., Xia, Y., et al. (2015). Reduced ABA accumulation in the root system is caused by ABA exudation in upland rice (*Oryza sativa* L. var. Gaoshan1) and this enhanced drought adaptation. *Plant Cell Physiol.* 56, 951–964. doi: 10.1093/pcp/pcv022

Smet, D., Opdebeeck, H., and Vandepoele, K. (2023). Predicting transcriptional responses to heat and drought stress from genomic features using a machine learning approach in rice. *Front. Plant Sci.* 14:1212073. doi: 10.3389/fpls.2023.1212073

Sun, Y., Wang, C., Chen, H. Y., and Ruan, H. (2020). Response of plants to water stress: a meta-analysis. *Front. Plant Sci.* 11:978. doi: 10.3389/fpls.2020.00978

Waititu, J. K., Zhang, X., Chen, T., Zhang, C., Zhao, Y., and Wang, H. (2021). Transcriptome analysis of tolerant and susceptible maize genotypes reveals novel insights about the molecular mechanisms underlying drought responses in leaves. *Int. J. Mol. Sci.* 22:6980. doi: 10.3390/ijms22136980

Wang, Y., Zhang, X., Huang, G., Feng, F., Liu, X., Guo, R., et al. (2019). iTRAQ-based quantitative analysis of responsive proteins under PEG-induced drought stress in wheat leaves. *Int. J. Mol. Sci.* 20:2621. doi: 10.3390/ijms20112621

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Yin, M., Zheng, Z., Zhang, Y., Wang, S., Zuo, L., Lei, Y., et al. (2024). Identification of key genes and pathways for anaerobic germination tolerance in rice using weighted gene co-expression network analysis (WGCNA) in association with quantitative trait locus (QTL) mapping. *Rice* 17:37. doi: 10.1186/s12284-024-00714-y

Yue, J., Liu, J., Tang, W., Wu, Y. Q., Tang, X., Li, W., et al. (2020). Kiwifruit genome database (KGD): a comprehensive resource for kiwifruit genomics. *Horticult. Res.* 7:117. doi: 10.1038/s41438-020-0338-9

Zhang, N., Yin, Y., Liu, X., Tong, S., Xing, J., Zhang, Y., et al. (2017). The E3 ligase TaSAP5 alters drought stress responses by promoting the degradation of DRIP proteins. *Plant Physiol.* 175, 1878–1892. doi: 10.1104/pp.17.01319

Zhou, P., Enders, T. A., Myers, Z. A., Magnusson, E., Crisp, P. A., Noshay, J. M., et al. (2022). Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *Plant Cell* 34, 514–534. doi: 10.1093/plcell/koab267

Zhou, Y., Li, H., Chen, H., Yang, X., Yu, T., Wang, Y., et al. (2022). Proteomic investigation of molecular mechanisms in response to PEG-induced drought stress in soybean roots. *Plan. Theory* 11:1173. doi: 10.3390/plants11091173

Zhu, C., Zhao, L., Zhao, S., Niu, X., Li, L., Gao, H., et al. (2024). Utilizing machine learning and bioinformatics analysis to identify drought-responsive genes affecting yield in foxtail millet. *Int. J. Biol. Macromol.* 277:134288. doi: 10.1016/j.ijbiomac.2024.134288

Zhu, M., Xie, H., Wei, X., Dossa, K., Yu, Y., Hui, S., et al. (2019). WGCNA analysis of salt-responsive core transcriptome identifies novel hub genes in rice. *Genes* 10:719. doi: 10.3390/genes10090719

Zrimec, J., Börlin, C. S., Buric, F., Muhammad, A. S., Chen, R., Siewers, V., et al. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* 11:6141. doi: 10.1038/s41467-020-19921-4

Glossary

RNA-Seq - RNA sequencing

DEGs - Differentially Expressed Genes

RF-Boruta - Random Forest-Boruta

AUC - Area Under Curve

GO - Gene Ontology

WGCNA - Weighted Gene Co-expression Network Analysis

ML - Machine learning

SVM - Support Vector Machine

KNN - K-Nearest Neighbor

DT - Decision Tree

RF - Random Forest

NB - Naive Bayes

GLM - Generalized Linear Model

CNCB - China National Center for Bioinformation

NCBI - National Center for Biotechnology Information

STAR - Spliced Transcripts Alignment to a Reference

RSEM - RNA-Seq by Expectation Maximization

PCA - principal component analysis

MA - Minus-versus-Add

SSE - sum of squared error

WSS - within-cluster sum of squares

ROC - Receiver Operating Characteristic

BP - biological processes

MF - molecular functions

FPKM - Fragments Per Kilobase of transcript per Million fragments mapped

KEGG - Kyoto Encyclopedia of Genes and Genomes

TF - transcription factors

BAM - Binary Alignment/Map format