Supplementary

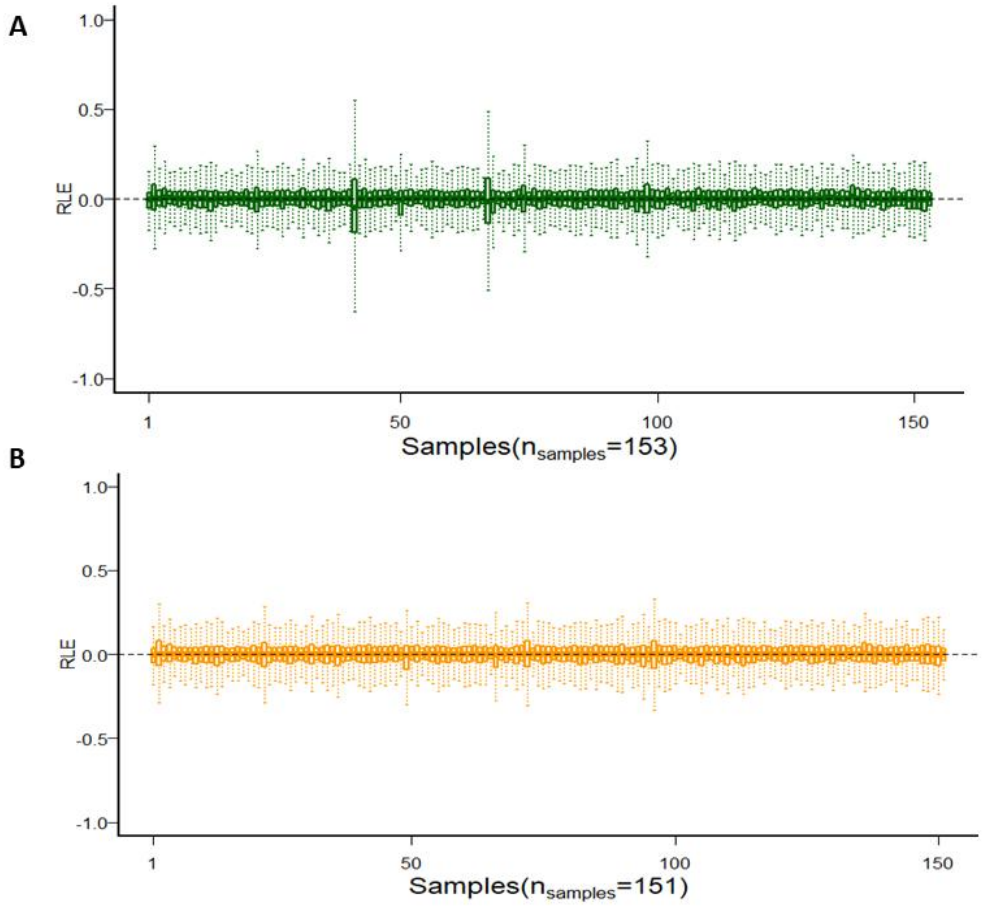

Figure S1. Relative Log Expression (RLE) plots before and after sample quality control

Each vertical box represents one of the 153 colorectal tumor samples. The y-axis shows per-gene RLE values ($\log_2$ expression minus the gene-wise median across samples), and the horizontal dashed line marks 0. Medians centered near 0 with similar interquartile ranges across samples indicate no global expression bias and comparable dispersion between samples.

(A) RLE plot of 153 tumor samples after DESeq2 normalization/VST. (B) RLE plot after excluding two outlier samples whose median RLE values deviated from 0. After outlier removal, the normalized RNA-seq data exhibited stable and unbiased expression distributions across samples, confirming successful normalization for downstream analyses.

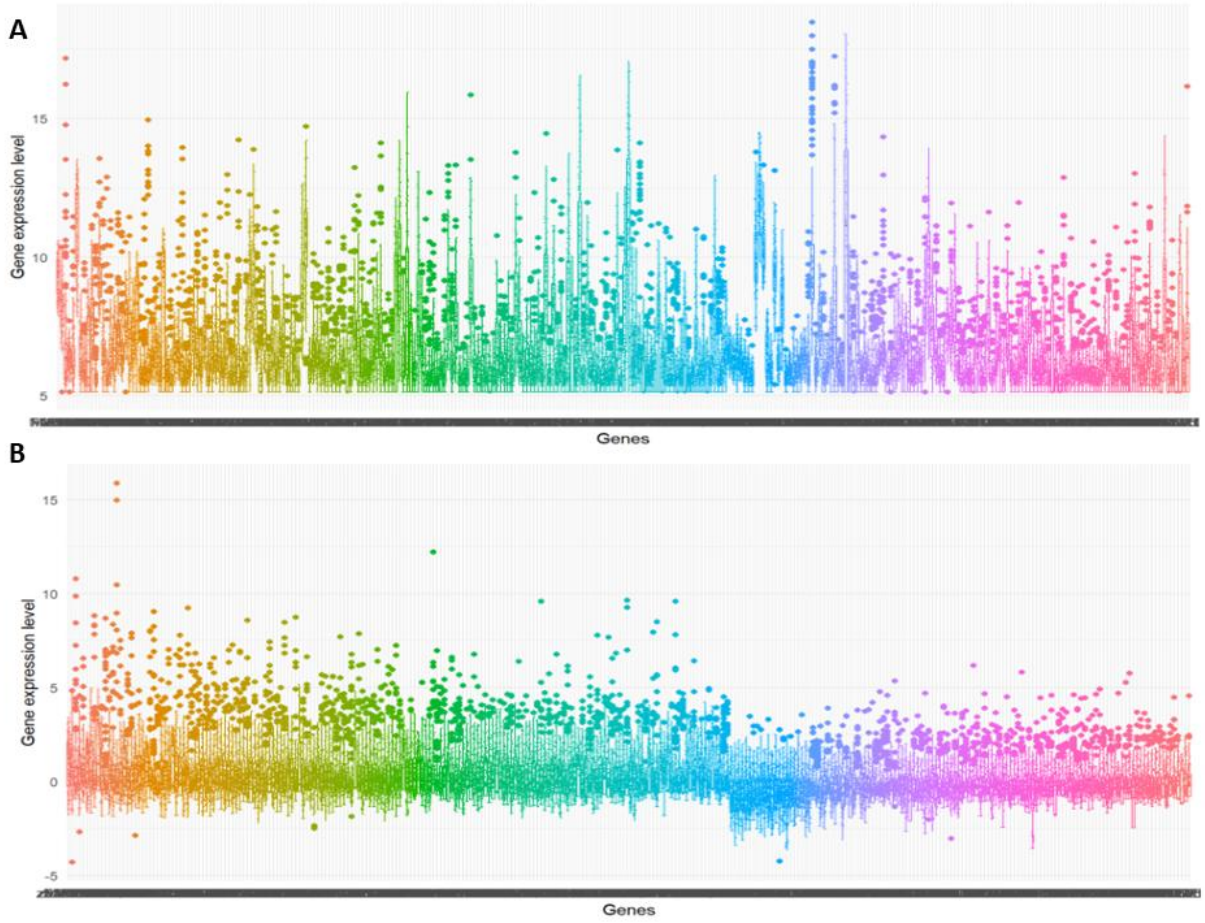

Figure S2. Transformation and z-score standardization of gene expression

(A) Distribution of gene expression values across all genes after DESeq2 normalization and variance-stabilizing transformation (VST). (B) The same expression matrix after per-gene z-score standardization, which centers each gene around zero and scales the variance uniformly across genes. This transformation facilitates comparison between genes and ensures compatibility for multivariate analyses and modeling.

For downstream visualization and predictive modeling, gene expression data were standardized to z-scores per gene. When specified, the mean and standard deviation (SD) were estimated from the lymph node–negative reference group to enable relative scaling across samples. The z-score transformation was computed as follows:

$$z(Gene_i) = \frac{Expression\ level_i - Mean_{lymph\ node\ negative}}{SD_{lymph\ node\ negative}}$$

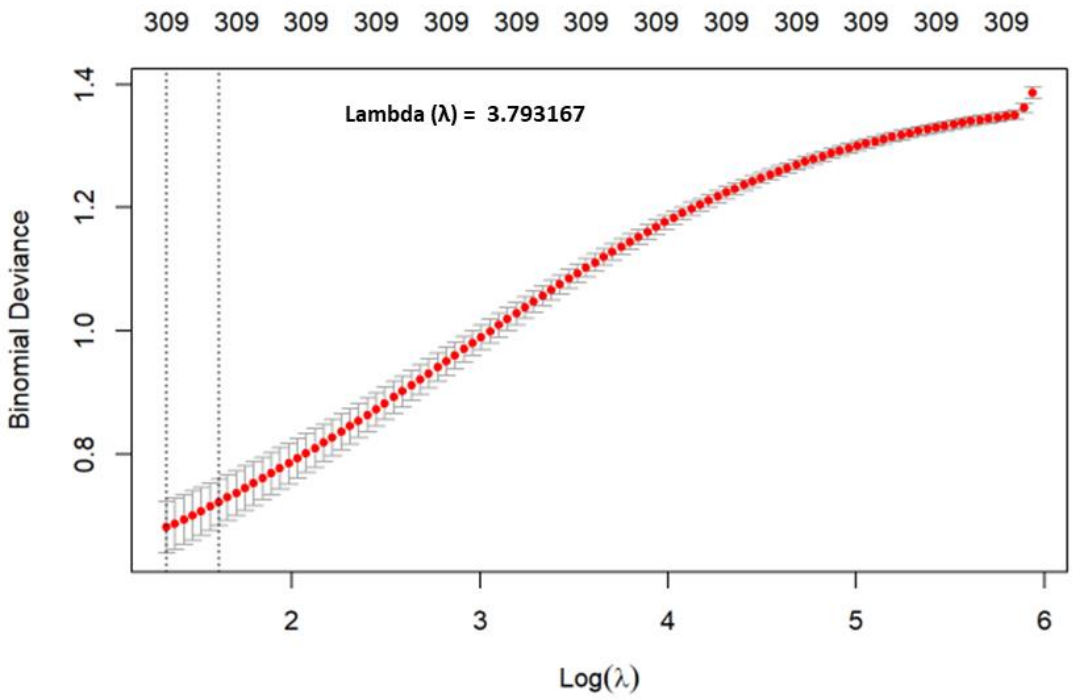

Figure S3. Ten-fold cross-validation for the Ridge regression model

Ten-fold cross-validation curve showing mean binomial deviance (red points) ± 1 SE (gray bars) across log(λ) values. Vertical dotted lines mark λ_min (left) and λ_1SE (right). The selected penalty at λ_min = 3.793167 minimizes deviance. Numbers along the top indicate the number of non-zero coefficients in the model at each λ.

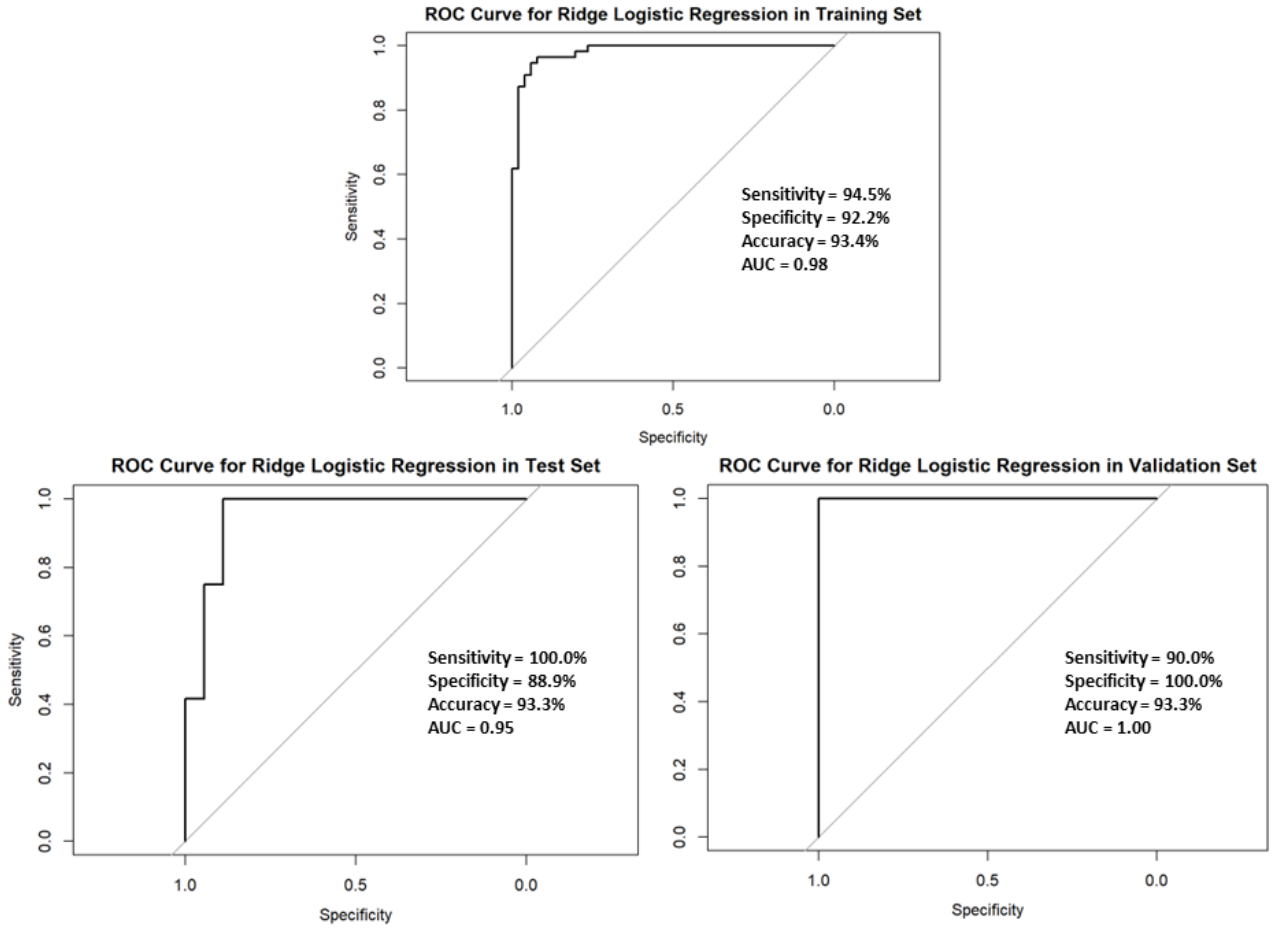

Figure S4: ROC curves for the 302 genes + 7 clinical variables for the Ride classifier

Receiver operating characteristic (ROC) curves for the Ridge logistic regression model incorporating 302 gene features and 7 clinical variables (sex, age, primary tumor location, T stage, lymphovascular invasion (LVI), perineural invasion (PNI), and tumor differentiation). The model was selected at $\lambda\_min = 3.793167$ and evaluated across three data partitions: training (top), test (bottom left), and hold-out validation (bottom right) sets. Performance metrics, including the area under the ROC curve (AUC), are displayed within each panel. Complete lists of model coefficients (β) are provided in Table S2.

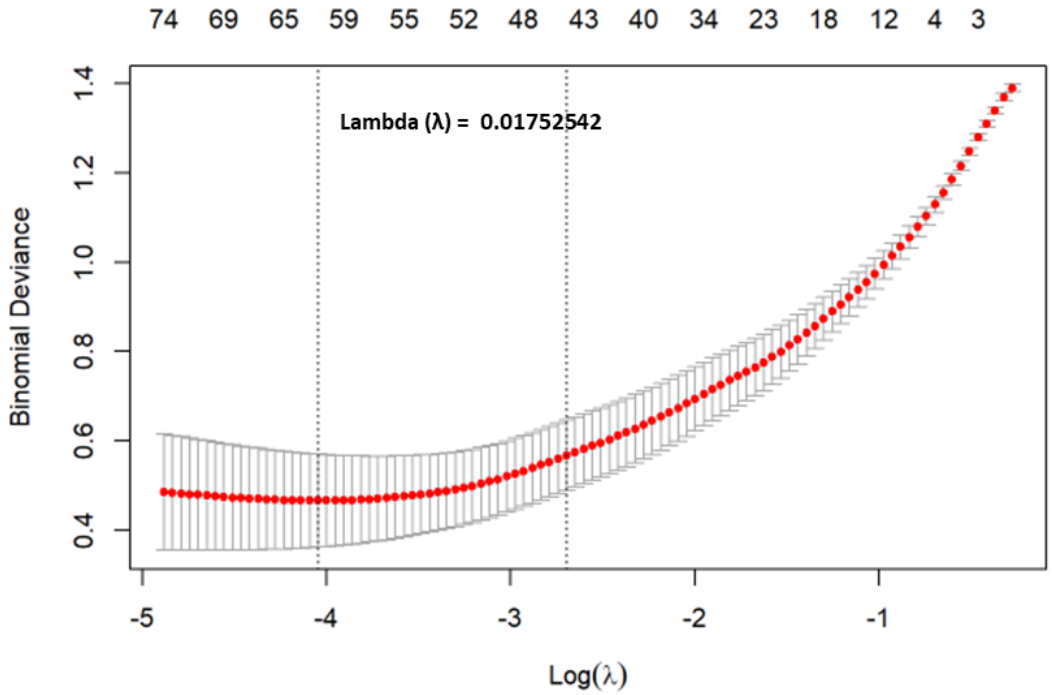

Figure S5. Ten-fold cross-validation for the Elastic Net regression model

Ten-fold cross-validation curve showing mean binomial deviance (red points) ± 1 SE (gray bars) across log(λ) values. Vertical dotted lines mark λ_min (left) and λ_1SE (right). The selected penalty at λ_min = 0.01752542 minimizes deviance. Numbers along the top indicate the number of non-zero coefficients in the model at each λ.

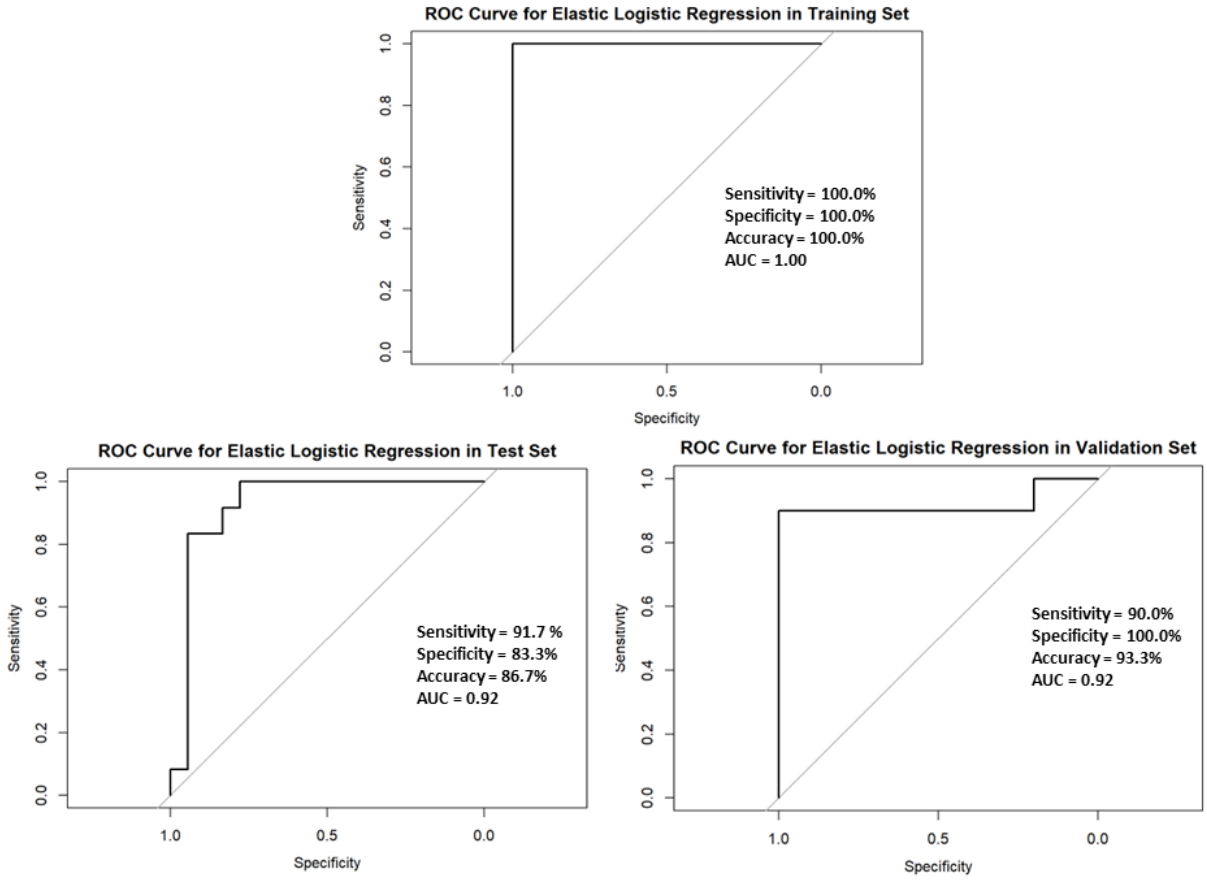

Figure S6: ROC curves for the 62 genes + 2 clinical variables for the Elastic classifier

Receiver operating characteristic (ROC) curves for the Elastic Net logistic regression model incorporating 62 gene features and two clinical variables (lymphovascular invasion [LVI] and tumor differentiation). The model was selected at $\lambda\_min = 0.0175$ and evaluated across three data partitions: training (top), test (bottom left), and hold-out validation (bottom right) sets. Performance metrics, including the area under the ROC curve (AUC), are displayed within each panel. Complete lists of model coefficients (β) are provided in Table S3.