# Application of AI-based Machine Learning Integrating Pathomics and Transcriptomics for Diagnosis and Prognostic Prediction in Clear Cell Renal Cell Carcinoma

**Keywords**
Artificial Intelligence , Deep Learning, Transcriptomics, Clear Cell Renal Cell Carcinoma, Pathological Omics, Prognostic Prediction

**Abstract**

Introduction
Clear cell renal cell carcinoma (ccRCC) is the most common and aggressive subtype of renal carcinoma, characterized by high molecular heterogeneity and variable clinical outcomes. Conventional prognostic models often lack the resolution to adequately stratify patient risk, underscoring the need for integrative, data-driven approaches.

Material and methods
We developed a novel diagnostic and prognostic model for ccRCC using an artificial intelligence-based machine learning approach (random forest algorithm) integrating pathomics and transcriptomic data. Whole-slide histopathological images and transcriptomic data were obtained from the TCGA-KIRC cohort. Radiomic features were extracted from histological slides, and gene modules related to tumor progression were identified using Weighted Gene Co-Expression Network Analysis (WGCNA). These modalities, together with clinical variables, were incorporated into a custom machine learning architecture to predict patient survival risk.

Results
The integrative model demonstrated strong predictive performance, effectively stratifying patients into high- and low-risk groups with significant differences in overall survival ($p < 0.001$). Functional enrichment analysis revealed that prognostic gene modules were associated with immune regulation, angiogenesis, and cell cycle pathways, highlighting their relevance in ccRCC pathogenesis.

Conclusions
This study presents a novel, AI-driven framework that combines multi-omics and imaging data to improve prognostic accuracy in ccRCC. The model offers potential utility in clinical decision-making and personalized treatment strategies, and may serve as a foundation for future precision oncology applications.

# Application of AI-based Machine Learning Integrating Pathomics and Transcriptomics for Diagnosis and Prognostic Prediction in Clear Cell Renal Cell Carcinoma

**Running Title:** Transcriptomics to Histopathology Integration

# Abstract

**Introduction:** Clear cell renal cell carcinoma (ccRCC) is the most common and aggressive subtype of renal carcinoma, characterized by high molecular heterogeneity and variable clinical outcomes. Conventional prognostic models often lack the resolution to adequately stratify patient risk, underscoring the need for integrative, data-driven approaches.

**Material and methods:** We developed a novel diagnostic and prognostic model for ccRCC using an artificial intelligence-based machine learning approach (random forest algorithm) integrating pathomics and transcriptomic data. Whole-slide histopathological images and transcriptomic data were obtained from the TCGA-KIRC cohort. Radiomic features were extracted from histological slides, and gene modules related to tumor progression were identified using Weighted Gene Co-Expression Network Analysis (WGCNA). These modalities, together with clinical variables, were incorporated into a custom machine learning architecture to predict patient survival risk.

**Results:** The integrative model demonstrated strong predictive performance, effectively stratifying patients into high- and low-risk groups with significant differences in overall survival ($p < 0.001$). Functional enrichment analysis revealed that prognostic gene modules were associated with immune regulation, angiogenesis, and cell cycle pathways, highlighting their relevance in ccRCC pathogenesis.

**Conclusions:** This study presents a novel, AI-driven framework that combines multi-omics and imaging data to improve prognostic accuracy in ccRCC. The model offers potential utility in clinical decision-making and personalized treatment strategies, and may serve as a foundation for future precision oncology applications.



**Keywords:** Clear Cell Renal Cell Carcinoma; Machine Learning; Pathological Omics; Transcriptomics; Prognostic Prediction; Artificial Intelligence

## Introduction

Clear cell renal cell carcinoma (ccRCC) is the most common subtype of renal cell carcinoma (RCC), accounting for 70%-80% of all RCC cases [1]. Although surgery is the primary treatment for early-stage ccRCC, nephrectomy, nephron-sparing surgery, and minimally invasive approaches have demonstrated excellent outcomes, achieving cancer-specific survival rates exceeding 90% in localized disease [2-4]. However, approximately 30% of patients experience recurrence and metastasis after surgery, leading to poor prognosis [5, 6]. Additionally, ccRCC is usually insensitive to chemotherapy and radiotherapy, making early diagnosis and accurate prognostic prediction crucial [7]. Traditional prognostic assessments mainly rely on clinical and pathological characteristics such as tumor size, stage, and grade [8], but these single clinical features often fail to capture the complex biological nature of the disease, limiting the accuracy of prognostic evaluations [9]. Therefore, developing more precise prognostic prediction tools for ccRCC through the integration of multidimensional data has become a critical challenge in both clinical and research settings.

In recent years, with the advancement of high-throughput sequencing technology, transcriptomics data have provided new insights into the molecular mechanisms of ccRCC. Multiple studies have shown that the development and progression of ccRCC are closely related to various gene mutations, abnormal expression, and regulatory networks [10]. For instance, the inactivation of the VHL gene is considered one of the primary driver events in ccRCC [11], affecting the activity of the HIF signaling pathway [12], which promotes tumor angiogenesis and cell proliferation. Additionally, abnormal expression of the TP53 gene has been associated with uncontrolled tumor cell proliferation and increased tumor invasiveness [13]. Thus, transcriptomics-based prognostic models can uncover molecular markers closely linked to ccRCC prognosis, providing potential therapeutic targets for personalized treatment [14, 15]. Beyond molecular-level research, radiomics has also emerged as a promising tool in ccRCC diagnosis and prognostic prediction [16]. Radiomics, by extracting quantitative features from imaging data—such

as texture, shape, and grayscale features—can capture tumor heterogeneity, microenvironment, and biological behavior [17, 18]. These quantitative imaging features provide more comprehensive tumor information compared to traditional imaging assessment methods, such as tumor size and morphology, thus offering new insights for the precise diagnosis and prognostic evaluation of ccRCC [19, 20]. However, challenges remain in the application of radiomics, such as the weak correlation between imaging features and clinical characteristics and the need to improve model generalizability. Therefore, combining radiomics with transcriptomics data through multi-omics integrated analysis may offer a more comprehensive and accurate tool for prognostic prediction in ccRCC.

The core rationale of this study lies in the fact that single clinical characteristics or molecular markers are insufficient to fully capture the complexity of tumors. By integrating pathological omics and transcriptomics data, alongside utilizing artificial intelligence and machine learning techniques, we aim to identify key features related to ccRCC prognosis, thereby enhancing the accuracy of predictive models. Furthermore, public databases such as The Cancer Genome Atlas - Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) provide high-quality datasets that include imaging, gene expression, and clinical information from ccRCC patients, enabling multi-omics integrated analysis. By developing a machine learning model based on radiomics and transcriptomics data, this study not only improves the prognostic prediction capability for ccRCC but also provides clinicians with a basis for personalized treatment, advancing precision medicine in oncology. In conclusion, this project is closely aligned with the clinical needs and research frontiers of ccRCC. Through the integration of multi-omics data and artificial intelligence technologies, it offers an innovative framework and technical pathway for ccRCC diagnosis, prognostic prediction, and personalized treatment.



# Materials and methods

### Collection of Clinical Data for ccRCC Patients

This study aims to develop an algorithmic model capable of accurately predicting the prognosis of ccRCC patients through the integrated analysis of transcriptomics, imaging, and clinical data. Relevant clinical data for ccRCC patients were collected from the TCGA-KIRC database. The dataset includes 103 ccRCC patients (77 males; median age 59 years, ranging from 34 to 79 years), of which 81 patients underwent CT scans, 19 underwent MRI scans, and 3 received both CT and MRI scans. Imaging features include tumor size (in mm), margins (clear or unclear boundaries), composition (solid or cystic), necrosis in solid tumors (categorized as 0%, 1%-33%, 34%-66%, or >66%), and calcification (present, absent, or indeterminate). Prognosis was determined based on tumor growth pattern (entirely endophytic, <50% exophytic, or ≥50% exophytic). The data used in this study were sourced from TCGA, which collects clinical data from various locations worldwide, ensuring both accuracy and reliability.

**Quality Control and Preprocessing of Imaging Data for ccRCC Patients**

After processing the gene data, obtaining high-quality imaging data is a crucial step in constructing the integrated analysis model. First, pixel matrix values from the source DCM files were read and standardized to facilitate subsequent data processing, and a uniform naming convention was applied to all images. Based on this data, the study utilized Radiomics algorithms to extract features across eight different expression categories: first-order statistics, shape features, texture features, wavelet features, fractal features, morphological features, gray-level region matrix features, and image entropy features. A total of 116 features were extracted across these categories. To improve the accuracy of prognosis predictions, Lasso regression was applied to filter these features, aiding in the subsequent integration with gene data for further analysis.

**Quality Control and Preprocessing of Clinical Data for ccRCC Patients**

For the integrated analysis, this study performed basic preprocessing of the clinical data. Several key features were normalized, and the growth pattern of ccRCC (Growth Pattern) was used

as the standard for prognosis. "Entirely Endophytic" (Class 0) indicates that the tumor is fully endophytic, associated with a better prognosis; "<50% Exophytic" (Class 1) signifies that the majority of the tumor is endophytic, indicating an intermediate prognosis; "=>50% Exophytic" (Class 2) suggests that most of the tumor is exophytic, representing more severe disease and a poor prognosis. Other clinical information, such as age, sex, and basic tumor characteristics, was included as part of the clinical dataset. The visualization of the clinical features for 103 patients is shown in Figure 1.

**Quality Control and Preprocessing of Transcriptomics Data for ccRCC Patients**

The TCGA-KIRC dataset includes over 60,000 gene features per patient. To improve the accuracy of the prognostic model, this study first normalized the expression values of each gene using Z-scores, adjusting the values to fall within the range of [-5, 5]. The visualization of gene data for 103 patients is shown in Figure 2.

To reduce redundant features, a differential analysis was performed, resulting in the selection of 20,000 key gene features. Based on this, the Weighted Gene Co-expression Network Analysis (WGCNA) method was applied for feature analysis and data quality control, ensuring the integrity and quality of the data to provide a high-quality foundation for training the prognostic model, thereby enhancing its accuracy and reliability. After identifying the most important feature modules through WGCNA, a correlation analysis was conducted between gene expression and the previously extracted imaging and clinical data features to identify the most critical gene features. The workflow for obtaining optimal gene features is illustrated in Figure 3.

**Dataset Splitting and Model Validation Using Cross-Validation**

To ensure that the constructed model maintains strong predictive performance on unseen data and effectively prevents overfitting, the dataset was divided into training, validation, and testing sets. A random sampling method was used, with the sample ratio set at 8:1:1 for the three datasets.

Specifically, 80% of the total samples were randomly selected as the training set for model training and parameter adjustment. Then, 10% of the remaining 20% of samples were randomly selected as the validation set to assess model performance and optimize hyperparameters. Finally, the remaining 10% of the samples served as the test set for the final evaluation of the model's performance.

In addition, cross-validation was employed during model training to assess performance. Specifically, the dataset was divided into 10 subsets, with one subset serving as the validation set and the remaining nine as the training set. This process was repeated 10 times, each time using a different subset as the validation set. This method ensures that every data point is used for validation once, with the final result being the average of all validation outcomes, providing a more stable and reliable model evaluation. To ensure that other researchers can replicate the results of this study, all data processing and model training parameters, code implementations, and hyperparameter settings are documented and made public. The data set partition, pre-processing steps, model training process and hyperparameter setting are provided in the form of documents to ensure the repeatability and transparency of the research. Through this arrangement, other researchers can conduct experiments under the same conditions, thus verifying the results and conclusions of this study.



**Construction of the Integrated Prognostic Model for ccRCC**

In building the integrated prognostic model for ccRCC, the random forest was chosen as the base model due to its advantages in feature fusion and joint analysis. The prognostic model constructed in this study (Figure 4) ranks the importance of each feature based on its influence across all decision trees, identifying the most impactful features on model predictions and effectively analyzing feature interactions to enhance predictive power. Additionally, the model's hyperparameters were fine-tuned to achieve optimal performance. Because the model uses multiple decision trees, it exhibits strong robustness against noisy data and outliers. Moreover, its inherent

parallelization allows for simultaneous training of multiple decision trees, improving model training efficiency. In summary, random forest enhances model performance by constructing multiple decision trees, training on random samples and feature subsets, effectively analyzing feature importance and interactions, and improving noise resistance and training efficiency, ultimately boosting the model's predictive accuracy.

**Statistical Analysis**

In this study, we employed various software tools and statistical methods to construct and evaluate predictive models for surgical outcomes and postoperative recovery. Programming and data processing were primarily conducted in Jupyter Notebook, using Python for coding, and Visio for visualization. Data processing and visualization were performed using the pandas and Matplotlib libraries in Python. Statistical methods included data normalization, calculation of model accuracy, sensitivity, and specificity, as well as performance evaluation through ROC curves and confusion matrices.

During data processing, imaging and clinical feature data were normalized to eliminate scale differences between different features. Simultaneously, gene data were normalized using Z-scores to ensure that each feature contributes equally during model training, thereby improving the stability and efficiency of the training process. After normalization, the data were converted into an array format to facilitate subsequent feature extraction and model training.

During the model evaluation phase, we calculated the model's accuracy, sensitivity, and specificity. These statistical metrics are primarily used to assess the model's performance in real-world applications. Accuracy reflects the model's overall predictive capability, sensitivity measures the model's ability to identify positive cases, and specificity evaluates its ability to recognize negative cases. Calculating these metrics provides a comprehensive understanding of the model's performance, ensuring its effectiveness in clinical applications.

Additionally, we utilized a confusion matrix to further evaluate model performance. The

confusion matrix provides detailed insights into the model's predictions across different categories, helping to identify how well the model performs in specific cases. The selection of these statistical methods is based on their widespread use and proven reliability in the medical field, offering robust validation of the model's effectiveness.

To visualize the data, we employed the Matplotlib package to generate intuitive displays of the arrays, showing the spatial characteristics of the samples and the model's performance. This visualization not only aids in understanding the data distribution but also provides a clear representation of the model's behavior under different conditions, offering valuable insights for future research and clinical applications. By combining these statistical methods with visualization techniques, we are able to more comprehensively validate the model's effectiveness and ensure its reliability in practical use.



# Results

### Feature Extraction and Visualization of Imaging Data for ccRCC Patients

In this study, the imaging data collected from ccRCC patents were first normalized. Using the open-source radiomics software package PyRadiomics, we extracted 116 features from the imaging data, covering eight key categories: 1. Morphological features, including tumor volume, maximum diameter, minimum diameter, long axis, short axis, surface area, volume ratio, and shape index. 2. Texture features, such as gray-level co-occurrence matrix (GLCM) features, gray-level run-length matrix (GLRLM) features, gray-level size zone matrix (GLSZM) features, and neighborhood gray-tone difference matrix (NGTDM) features. 3. Shape features, including sphericity, roundness, long axis ratio, and curvature. 4. Grayscale histogram features, including mean, standard deviation, skewness, kurtosis, and entropy. 5. Wavelet features, representing features at different scales obtained through wavelet transformations. 6. Model-based features, such as tumor growth pattern and boundary clarity. 7. Radiomics features derived from radiomics algorithms, including vascular density and perfusion rate. 8. Other features, including patient demographics such as age, sex, and

tumor staging.

To identify the most important features, this study employed the Lasso regression method. Lasso regression is a regularization technique that applies a penalty to the regression coefficients, effectively eliminating irrelevant features and enhancing the predictive power of the model. Through Lasso regression analysis, the research team ultimately selected 12 key features that are closely associated with patient prognosis. A heatmap illustrating the similarity between these selected features is shown in Figure 5.

**Feature Extraction and Visualization of Gene Data for ccRCC Patients**

This study utilized the WGCNA method [21, 22] to extract and visualize features from the gene data of ccRCC patients, revealing relevant gene regulatory networks. Initially, the research team integrated a dataset of 20,000 pre-filtered genes. The WGCNA algorithm was then applied to analyze these genes, constructing a gene co-expression network. WGCNA is a systems biology approach that groups genes with highly correlated expression into modules and performs functional enrichment analysis for each module, uncovering interactions between genes. Through WGCNA, the team identified 21 gene modules, each representing a group of highly correlated genes. To better understand the functions of these gene modules, the team conducted functional analyses (Figure 6) and presented the correlations between gene modules, functional enrichment results, and their associations with clinical characteristics.

Additionally, to further analyze each gene module, we visualized the topological overlap matrix between gene modules, as shown in Figure 7. The topological overlap matrix visualization depicts the similarity between gene modules by considering both the internal connections within a module and its connections to other modules, making it a more comprehensive reflection of inter-module relationships compared to simple correlation analysis. By performing hierarchical clustering on the topological overlap matrix, we grouped gene modules with high topological overlap into clusters. These clusters typically represent gene sets with common functions or

regulatory mechanisms, aiding in the understanding of gene functionality and interactions. Within each gene module, genes with high topological overlap are often considered key genes, playing significant roles in the module's functions. In summary, the topological overlap matrix offers a complete understanding of the relationships between gene modules, facilitating subsequent module identification, functional analysis, and key gene recognition and providing crucial insights into gene functions and interactions.

**Integrated Analysis of Clinical Data for ccRCC Patients**

To further analyze the clinical characteristics of patients, we conducted a correlation analysis between the gene modules, imaging data, and clinical features obtained in the study, as shown in Figure 8.

As seen in Figure 8A, Module 1 shows the highest similarity across all clinical features, with the largest expression levels. Additionally, histological grade (HistGrade) of ccRCC, stage of ccRCC, clinical axis length, and age are the most strongly correlated clinical features with the genes. Similarly, Figure 8B reveals that Module 1 also has the highest similarity and expression levels across all imaging features compared to other modules. In summary, the genes in Module 1 are most strongly associated with both clinical and imaging data, making them the primary focus for further analysis in the prognostic model training. This allows our model to concentrate on the most important features, potentially identifying target genes or pathways for drugs aimed at suppressing tumor growth and improving patient survival.

**Division and Significance of Clinical Datasets for ccRCC Patients**

After preprocessing, clinical data from a total of 103 patients were obtained. Of these, 9 patients were classified as having a favorable prognosis (Entirely Endophytic: Class 0), 49 were classified with an intermediate prognosis (<50% Exophytic: Class 1), and 45 were classified with a poor prognosis (=>50% Exophytic: Class 2). The dataset was then split into training, testing, and

validation sets in an 8:1:1 ratio. The distribution of data across the different sets is shown in Table 1.

The preprocessing and division of clinical data for ccRCC patients provide a solid foundation for subsequent model training and evaluation. With further research, we anticipate gaining a deeper understanding of the prognoses of ccRCC patients, thereby offering more effective guidance for clinical treatment.

**Results and Visualization of the Prognostic Model for ccRCC Based on Integrated Analysis**

After constructing the model, we compared the predicted results with the original prognosis categories by calculating precision, recall, and F1-score, as shown in Table 2. Due to the small sample size, the results of the validation set were combined with those of the test set.

As shown in the table, the model meets certain prognostic standards and is capable of providing relatively accurate prognostic predictions for ccRCC. Although the predictive performance for Class 0 is suboptimal, the prognostic accuracy for Classes 1 and 2 is satisfactory. It is important to note that Class 0 contains only nine patients, resulting in an extremely limited sample size and a markedly imbalanced class distribution. This substantially contributes to the reduced predictive performance observed for this group and represents a key aspect requiring further improvement and validation. To further illustrate the accuracy of the model, we visualized the confusion matrix, as shown in Figure 9.

Additionally, to analyze the highly expressed genes in ccRCC, we deconstructed the integrated prognostic model and identified the top eight genes contributing most to the model. The contribution of these genes is visualized in Figure 10.

From the analysis, it is evident that the eight genes with the highest weights are closely related to ccRCC. The TP53 gene, known as a tumor suppressor gene, plays a crucial role in maintaining genomic stability and preventing tumor development [23, 24]. It encodes the p53 protein, a transcription factor that responds to various stress signals. In this study, patients with poor prognosis exhibited negative expression of this gene, indicating that their TP53 gene could not adequately

respond to the stress signals associated with ccRCC, leading to dysregulation in processes such as cell cycle control, DNA repair, and apoptosis, ultimately allowing tumor growth and progression. The VHL gene is the most frequently mutated gene in ccRCC, and mutations in VHL result in the stabilization of the HIF-1α protein, which promotes angiogenesis and tumor growth [25, 26]. Studies have suggested that VHL mutations may suppress TP53 mutations, as the loss of VHL decreases cellular sensitivity to DNA damage, thus reducing the likelihood of TP53 mutations. These findings confirm that our prognostic model correctly identifies the most important genetic features, further demonstrating its strong predictive capability.

## Discussion

RCC is one of the most common malignancies globally, with ccRCC representing the majority of cases [27, 28]. However, existing clinical prognosis prediction methods are often limited by single-source data [29], making it difficult to comprehensively reflect the complex biological characteristics of tumors. Therefore, developing more integrated, multidimensional predictive models holds significant clinical and research importance. In this study, we developed an innovative prognostic prediction model for ccRCC based on an artificial intelligence-driven machine learning approach (random forest ensemble model) integrating pathomics and transcriptomic data. The model demonstrates significant research value and broad clinical application potential. As tumor diagnosis and prognosis techniques rapidly advance, traditional single-source data are increasingly insufficient for accurately diagnosing complex diseases [30]. By integrating multi-omics data and analyzing both imaging features and gene expression, this study greatly enhances the accuracy of ccRCC diagnosis. This interdisciplinary approach not only improves the model's ability to differentiate patient prognoses but also provides a more scientific basis for personalized treatment, promoting the integration of tumor research and clinical application.

In terms of data processing and model construction, the study applied Radiomics algorithms to extract up to 116 features from imaging data of ccRCC patients, including morphological, texture,

and grayscale histogram features. This rich imaging information provides a comprehensive description of the pathological phenotype of ccRCC. Simultaneously, WGCNA was used to identify key gene modules from extensive transcriptomics data, with these modules being closely associated with ccRCC prognosis, particularly the TP53 and VHL genes, whose mutations are linked to malignant biological behavior in ccRCC. The combined application of these imaging and genetic features in the model significantly enhances the accuracy of prognostic stratification for ccRCC patients, effectively distinguishing between groups with good, intermediate, and poor prognoses.

The most significant innovation of this study is the introduction of a prognostic prediction model based on random forest, which integrates multidimensional tumor features. The random forest model excels at handling complex datasets and interactions between multiple features. By constructing multiple decision trees and ranking feature importance, the model accurately identifies key features critical for prognostic prediction, such as tumor maximum diameter, boundary clarity, and gene expression levels, greatly improving both the predictive power and stability of the model [31-34]. Additionally, through cross-validation and testing set validation, the model demonstrates high accuracy, sensitivity, and specificity, indicating strong generalizability.

From a clinical perspective, this model offers a novel tool for personalized treatment and prognostic management of ccRCC patients. By applying the prognostic model, clinicians can assess patient prognosis at an earlier stage and develop personalized treatment plans based on the model's stratification results, thereby improving treatment efficacy and reducing unnecessary risks. For patients with poor prognosis, the model aids in identifying specific high-risk genes, providing critical theoretical support for the development and application of targeted therapies. Additionally, the findings of this study can guide postoperative follow-up and treatment adjustments, ultimately contributing to improved survival rates and quality of life for ccRCC patients.

It is important to emphasize that, in clinical practice, active surveillance is often a more appropriate management strategy for elderly patients with significant comorbidities whose tumors remain confined to the kidney. Numerous studies have demonstrated that, for localized renal cell

carcinoma, radical nephrectomy, nephron-sparing surgery, and minimally invasive therapies can all achieve cancer-specific survival rates exceeding 90% in the general population [2-4]. However, such therapeutic benefits may not extend to older patients with multiple comorbid conditions. In these individuals, mortality risk is more strongly driven by competing risks such as cardiovascular and metabolic diseases, and the risks associated with aggressive surgical treatment outweigh any potential oncologic benefit. Consequently, neither active intervention strategies nor the prognostic prediction system developed in this study have meaningful clinical applicability for this patient subgroup.

Despite its strengths, this study has several limitations that warrant cautious interpretation of the findings. First, the model was built on 103 patients from the TCGA-KIRC cohort, representing a relatively limited sample size. Notably, the favorable-prognosis group included only nine cases, resulting in pronounced class imbalance that may lead to unstable parameter estimation, increased risk of overfitting, and reduced predictive performance in low-risk patients. Second, the imaging data exhibited substantial heterogeneity, encompassing CT, MRI, and combined imaging modalities. Variability in imaging equipment, acquisition parameters, and contrast agent usage was not fully standardized, which may affect the stability and reproducibility of radiomic features and introduce potential bias. Furthermore, this study used tumor growth pattern as a surrogate endpoint for prognosis. Although it correlates with tumor aggressiveness and postoperative outcomes, it remains an indirect measure and cannot substitute for clinically meaningful endpoints such as overall survival or disease-free survival. Due to the lack of complete imaging-survival paired data in the TCGA imaging cohort, we were unable to construct survival-based prognostic models directly. In summary, this work should be regarded as an exploratory starting point that integrates multi-omics data with artificial intelligence methodologies. Future research should rely on multicenter, large-scale cohorts with unified acquisition of imaging and survival data, accompanied by standardized imaging workflows, expanded sample size, and model validation using OS/DFS endpoints. Such efforts will be essential for improving the clinical generalizability and translational

value of the proposed model.

In the future, with the continuous development of multi-omics data and machine learning algorithms, the prognostic model proposed in this study could be further optimized and applied more widely in clinical practice. Particularly in the era of personalized medicine, leveraging more precise tumor characterization and clinical stratification tools could drive advances in the precision diagnosis and treatment of ccRCC. Further multi-center large-scale studies and clinical trials will help expand the application of this model across diverse patient populations, ultimately contributing to improved survival rates and quality of life for ccRCC patients.

In conclusion, this study successfully developed a ccRCC prognostic prediction model by integrating imaging and genomic data, leveraging machine learning technologies. The model demonstrates both innovation and practical utility, offering new insights for future cancer research while advancing the application of precision medicine in the diagnosis and treatment of ccRCC.

# Abbreviations

ccRCC: Clear Cell Renal Cell Carcinoma

GLCM: Gray-Level Co-Occurrence Matrix

GLRLM: Gray-Level Run-Length Matrix

GLSZM: Gray-Level Size Zone Matrix

HistGrade: Histological Grade

NGTDM: Neighborhood Gray-Tone Difference Matrix

RCC: Renal Cell Carcinoma

TCGA-KIRC: The Cancer Genome Atlas - Kidney Renal Clear Cell Carcinoma

WGCNA: Weighted Gene Co-Expression Network Analysis

## Ethical statement

Not applicable.

## Acknowledgment

Not applicable.

## Conflict of interest

The authors declare no conflict of interest.

## Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

# References


1. Wolf MM, Kimryn Rathmell W, Beckermann KE. Modeling clear cell renal cell carcinoma and therapeutic implications. *Oncogene*. 2020;39(17):3413-3426. doi:10.1038/s41388-020-1234-3

2. Ciccarese C, Strusi A, Arduini D, et al. Post nephrectomy management of localized renal cell carcinoma. From risk stratification to therapeutic evidence in an evolving clinical scenario. *Cancer Treatment Reviews*. 2023;115:102528. doi:10.1016/j.ctrv.2023.102528

3. Bai H, Jiang W, Wang D, Shou J, Li C, Xing N. Efficacy and safety of surgery in renal carcinoma patients 75 years and older: a retrospective analysis. *BMC Urol*. 2022;22(1). doi:10.1186/s12894-022-01088-3

4. Assessment of the effectiveness of radiofrequency ablation as a technique for destroying small renal tumors in patients older than 70. *Cent european J Urol*. Published online 2020. doi:10.5173/ceju.0310

5. Majdoub M, Yanagisawa T, Quhal F, et al. Role of clinicopathological variables in predicting recurrence and survival outcomes after surgery for non-metastatic renal cell carcinoma: Systematic review and meta-analysis. *Intl Journal of Cancer*. 2023;154(7):1309-1323. doi:10.1002/ijc.34793

6. Mattila KE, Vainio P, Jaakkola PM. Prognostic Factors for Localized Clear Cell Renal Cell Carcinoma and Their Application in Adjuvant Therapy. *Cancers*. 2022;14(1):239. doi:10.3390/cancers14010239

7. Huang G, Liao J, Cai S, et al. Development and validation of a prognostic nomogram for predicting cancer-specific survival in patients with metastatic clear cell renal carcinoma: A study based on SEER database. *Front Oncol*. 2022;12. doi:10.3389/fonc.2022.949058

8. Taneja K, Williamson SR. Updates in Pathologic Staging and Histologic Grading of Renal Cell Carcinoma. *Surgical Pathology Clinics*. 2018;11(4):797-812. doi:10.1016/j.path.2018.07.004

9. Delahunt B, Eble JN, Egevad L, Samaratunga H. Grading of renal cell carcinoma. *Histopathology*. 2018;74(1):4-17. doi:10.1111/his.13735

10. Bai S, Wu Y, Yan Y, et al. Construct a circRNA/miRNA/mRNA regulatory network to

explore potential pathogenesis and therapy options of clear cell renal cell carcinoma. *Sci Rep*. 2020;10(1). doi:10.1038/s41598-020-70484-2

11. Hsieh JJ, Le VH, Oyama T, Ricketts CJ, Ho TH, Cheng EH. Chromosome 3p Loss–Orchestrated VHL, HIF, and Epigenetic Deregulation in Clear Cell Renal Cell Carcinoma. *JCO*. 2018;36(36):3533-3539. doi:10.1200/jco.2018.79.2549

12. Zhang X, Li S, He J, et al. TET2 Suppresses VHL Deficiency-Driven Clear Cell Renal Cell Carcinoma by Inhibiting HIF Signaling. *Cancer Research*. 2022;82(11):2097-2109. doi:10.1158/0008-5472.can-21-3013

13. Yang Y, Luo Y, Huang S, Tao Y, Li C, Wang C. MKRN1/2 serve as tumor suppressors in renal clear cell carcinoma by regulating the expression of p53. *CBM*. 2023;36(4):267-278. doi:10.3233/cbm-210559

14. Zhang J, Zhang X, Piao C, et al. A long non-coding RNA signature to improve prognostic prediction in clear cell renal cell carcinoma. *Biomedicine & Pharmacotherapy*. 2019;118:109079. doi:10.1016/j.biopha.2019.109079

15. Luo J, Xie Y, Zheng Y, et al. Comprehensive insights on pivotal prognostic signature involved in clear cell renal cell carcinoma microenvironment using the ESTIMATE algorithm. *Cancer Medicine*. 2020;9(12):4310-4323. doi:10.1002/cam4.2983

16. Yi X, Xiao Q, Zeng F, et al. Computed Tomography Radiomics for Predicting Pathological Grade of Renal Cell Carcinoma. *Front Oncol*. 2021;10. doi:10.3389/fonc.2020.570396

17. Dwivedi DK, Xi Y, Kapur P, et al. Magnetic Resonance Imaging Radiomics Analyses for Prediction of High-Grade Histology and Necrosis in Clear Cell Renal Cell Carcinoma: Preliminary Experience. *Clinical Genitourinary Cancer*. 2021;19(1):12-21.e1. doi:10.1016/j.clgc.2020.05.011

18. Liu DH, Dani KA, Reddy SS, et al. Radiogenomic Associations Clear Cell Renal Cell Carcinoma: An Exploratory Study. *Oncology*. 2023;101(6):375-388. doi:10.1159/000530719

19. Gao J, Ye F, Han F, Jiang H, Zhang J. A radiogenomics biomarker based on immunological heterogeneity for non-invasive prognosis of renal clear cell carcinoma. *Front Immunol*. 2022;13.

doi:10.3389/fimmu.2022.956679

20. Choi JW, Hu R, Zhao Y, et al. Preoperative prediction of the stage, size, grade, and necrosis score in clear cell renal cell carcinoma using MRI-based radiomics. *Abdom Radiol*. 2021;46(6):2656-2664. doi:10.1007/s00261-020-02876-x

21. Petrosyan V, Dobrolecki LE, Thistlethwaite L, et al. Identifying biomarkers of differential chemotherapy response in TNBC patient-derived xenografts with a CTD/WGCNA approach. *iScience*. 2023;26(1):105799. doi:10.1016/j.isci.2022.105799

22. Liu Y. CWGCNA: an R package to perform causal inference from the WGCNA framework. *NAR Genomics and Bioinformatics*. 2024;6(2). doi:10.1093/nargab/lqae042

23. Donehower LA, Soussi T, Korkut A, et al. Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. *Cell Reports*. 2019;28(5):1370-1384.e5. doi:10.1016/j.celrep.2019.07.001

24. Fito-Lopez B, Salvadores M, Alvarez MM, Supek F. Prevalence, causes and impact of TP53-loss phenocopying events in human tumors. *BMC Biol*. 2023;21(1). doi:10.1186/s12915-023-01595-1

25. Hu J, Tan P, Ishihara M, et al. Tumor heterogeneity in VHL drives metastasis in clear cell renal cell carcinoma. *Sig Transduct Target Ther*. 2023;8(1). doi:10.1038/s41392-023-01362-2

26. Kurlekar S, Lima JDCC, Li R, et al. Oncogenic Cell Tagging and Single-Cell Transcriptomics Reveal Cell Type–Specific and Time-Resolved Responses to Vhl Inactivation in the Kidney. *Cancer Research*. 2024;84(11):1799-1816. doi:10.1158/0008-5472.can-23-3248

27. Wu J, Zhang F, Zhang J, et al. A Novel miRNA-Based Model Can Predict the Prognosis of Clear Cell Renal Cell Carcinoma. *Technol Cancer Res Treat*. 2021;20. doi:10.1177/15330338211027923

28. Shuiping Y, xu dandan, meng Z, et al. Kidney microbiome in patients with kidney carcinoma: Role of SA and SNZ gene expression. *Arch Med Sci*. Published online October 29, 2021.

doi:10.5114/aoms/143148

29. Liu G, Xiong D, Che Z, Chen H, Jin W. A novel inflammation-associated prognostic signature for clear cell renal cell carcinoma. *Oncol Lett*. 2022;24(3). doi:10.3892/ol.2022.13427

30. Glennon KI, Maralani M, Abdian N, et al. Rational Development of Liquid Biopsy Analysis in Renal Cell Carcinoma. *Cancers*. 2021;13(22):5825. doi:10.3390/cancers13225825

31. Li J, Tian Y, Zhu Y, et al. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. *Artificial Intelligence in Medicine*. 2020;103:101814. doi:10.1016/j.artmed.2020.101814

32. Zhao D, Kim DY, Chen P, et al. Pan-Cancer Survival Classification With Clinicopathological and Targeted Gene Expression Features. *Cancer Inform*. 2021;20. doi:10.1177/11769351211035137

33. Qiu X, Gao J, Yang J, et al. A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. *Front Oncol*. 2020;10. doi:10.3389/fonc.2020.551420

34. Watts J, Allen E, Mitoubsi A, et al. Adapting Random Forests to Predict Obesity-Associated Gene Expression. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Published online July 11, 2022:4407-4410. doi:10.1109/embc48229.2022.9871234

# Figure Legends

**Figure 1. Visualization of Clinical Feature Data.**

**Figure 2. Visualization of Gene Data.**

Note: The above figure visualizes gene expression values for 103 patients. Red indicates higher gene expression, while cyan indicates that the gene is not normally expressed and is negatively expressed.

**Figure 3. Workflow for Identifying Optimal Gene Features by Integrating Imaging and Clinical Features.**

Note: The WGCNA primarily involves six steps: calculating the direct correlation between genes (gray module in the figure), determining the soft threshold for gene modules (gray module), calculating the indirect correlation between genes (purple module), calculating the Euclidean distance between gene modules (blue module), obtaining the gene tree pathway between gene blocks (orange module), and dividing and cutting gene modules (yellow module). These methods are combined to classify all genes, facilitating the subsequent integration of imaging data and clinical features for analysis.

**Figure 4. Prognostic Model Based on Random Forest Integrating ccRCC Genes, Imaging, and Clinical Features.**

**Figure 5. Similarity Between Features Selected Through Lasso Regression.**

Note: The 12 selected features are as follows: Maximum 2D Diameter Row (maximum two-dimensional diameter of the tumor in the sagittal plane), Minor Axis Length (short axis length of the tumor), Maximum Diameter (maximum tumor diameter), Perimeter (tumor region perimeter), Pixel Surface (number of surface pixels in the tumor region), Gray-Level Non-Uniformity 1~3 (non-uniformity of gray level), Run Length Non-Uniformity (non-uniformity of running length), Size Zone Non-Uniformity (non-uniformity of area size), Dependence Non-Uniformity (dependent on non-uniformity), and Large Dependence Emphasis (gray-level dependence matrix). Each of these features is crucial for prognosis, with a significant impact factor.

**Figure 6. Visualization of Gene Analysis Based on WGCNA.**

Note: The WGCNA method transforms the direct correlation matrix into an indirect correlation matrix to calculate the soft threshold. Using this soft threshold, the original correlation network is converted into a scale-free network. (A) and (B) present the scale-free network in different coordinate systems. (A) the vertical axis shows the evaluation metric $r^2$ for the scale-free network. The closer $r^2$ is to 1, the more the network approaches a scale-free structure, with $r^2$ typically required to be greater than 0.8 or 0.9; (B) the vertical axis represents the mean connectivity, which decreases as the β value increases. By combining these two metrics, we typically select the β value where $r^2$ first reaches 0.8, 0.9, or higher. This β value is then used to convert the correlation matrix into an adjacency matrix, which represents the connections between nodes in the network and is key to building the scale-free network; (C) shows a dynamic tree cut method applied to the topological overlap matrix based on the scale-free network, revealing a co-expression network analysis where each gene module contains co-expressed genes that share similar gene characteristics.

**Figure 7. Visualization of the Topological Overlap Matrix Between Gene Modules.**

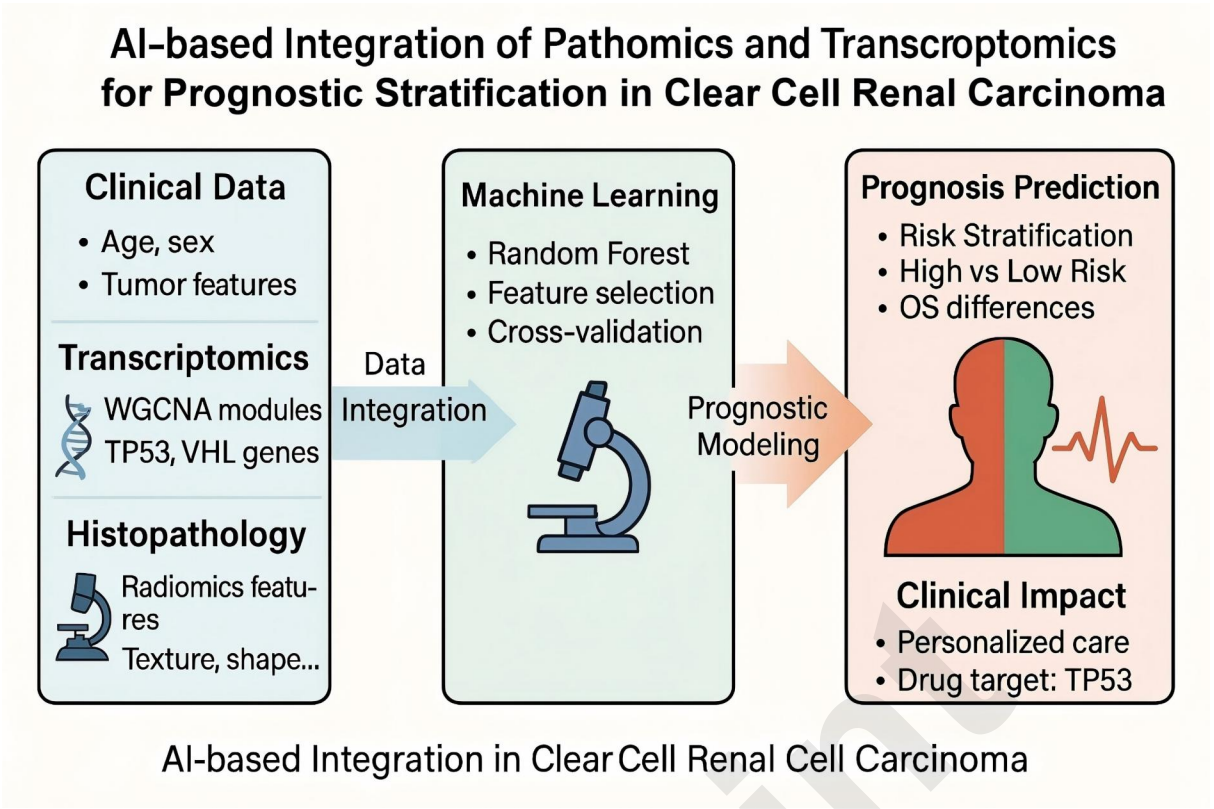
Note: The values in the topological overlap matrix range from 0 to 1, with higher values indicating greater similarity between gene modules. It is evident that smaller gene modules exhibit higher similarity, while larger gene modules show lower similarity. This suggests that smaller gene modules may share similar functions or regulatory mechanisms, whereas larger gene modules might have distinct functions or regulatory mechanisms.

**Figure 8. Evaluation of the Model.**

Note: (A) shows the correlation analysis between gene modules and clinical features; (B) shows the correlation analysis between gene modules and imaging data. In both figures, "Correlation" represents similarity, with darker red indicating greater similarity. The size of the circles represents the expression level of the gene modules for the given trait, with larger circles indicating higher expression levels.

**Figure 9. Visualization of the Confusion Matrix for Model Results.**

**Figure 10. Contribution of the Top Eight Genes with the Highest Weight in the Prognostic Model.**




AI-based Integration in Clear Cell Renal Cell Carcinoma

**Table 1. Sample distribution of data sets.**

| | **Class 0 (Entirely Endophytic)** | **Class 1 (<50% exophytic)** | **Class 2 (=>50% exophytic)** |
|---|---|---|---|
| Tarin Set | 6 | 40 | 36 |
| Test Set | 1 | 4 | 5 |
| Val Set | 2 | 5 | 4 |
| Total | 9 | 49 | 45 |

**Table 2. Results of the prognostic model.**

| Classification | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.67 | 0.67 | 3 |
| 1 | 0.75 | 0.67 | 0.71 | 9 |
| 2 | 0.80 | 0.89 | 0.84 | 9 |


Figure 1. Visualization of Clinical Feature Data.

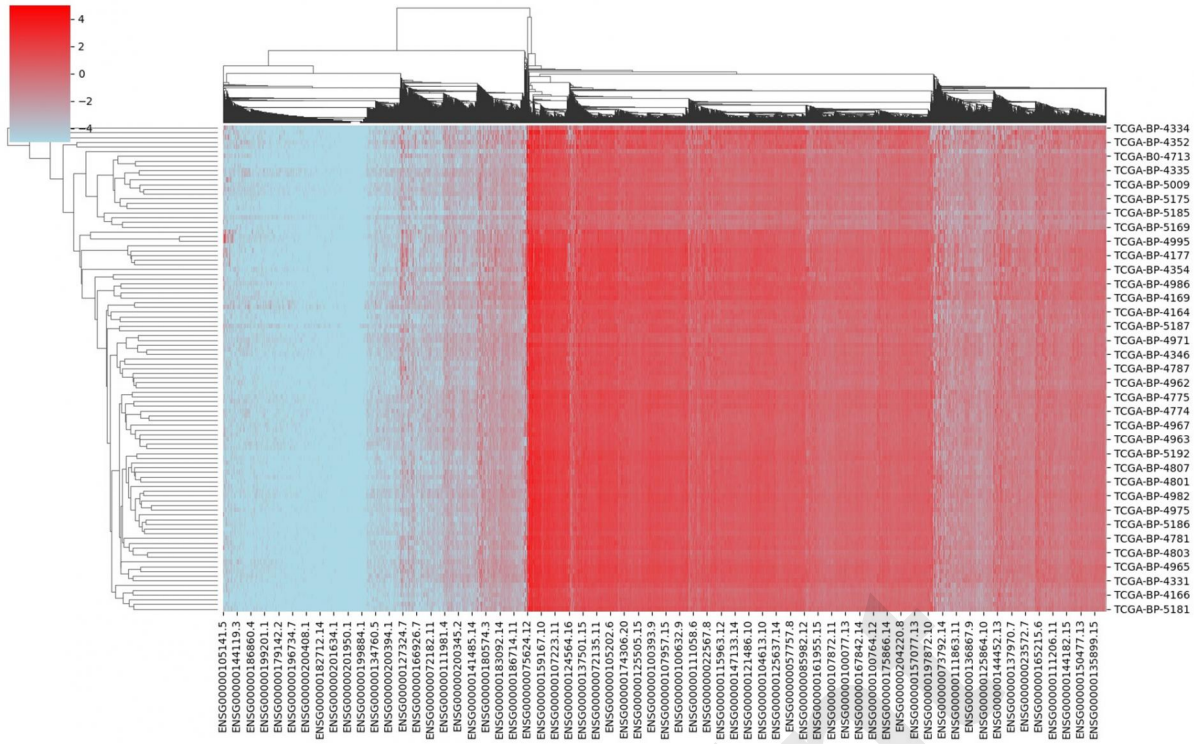



Figure 2. Visualization of Gene Data.
Note: The above figure visualizes gene expression values for 103 patients. Red indicates higher gene expression, while cyan indicates that the gene is not normally expressed and is negatively expressed.

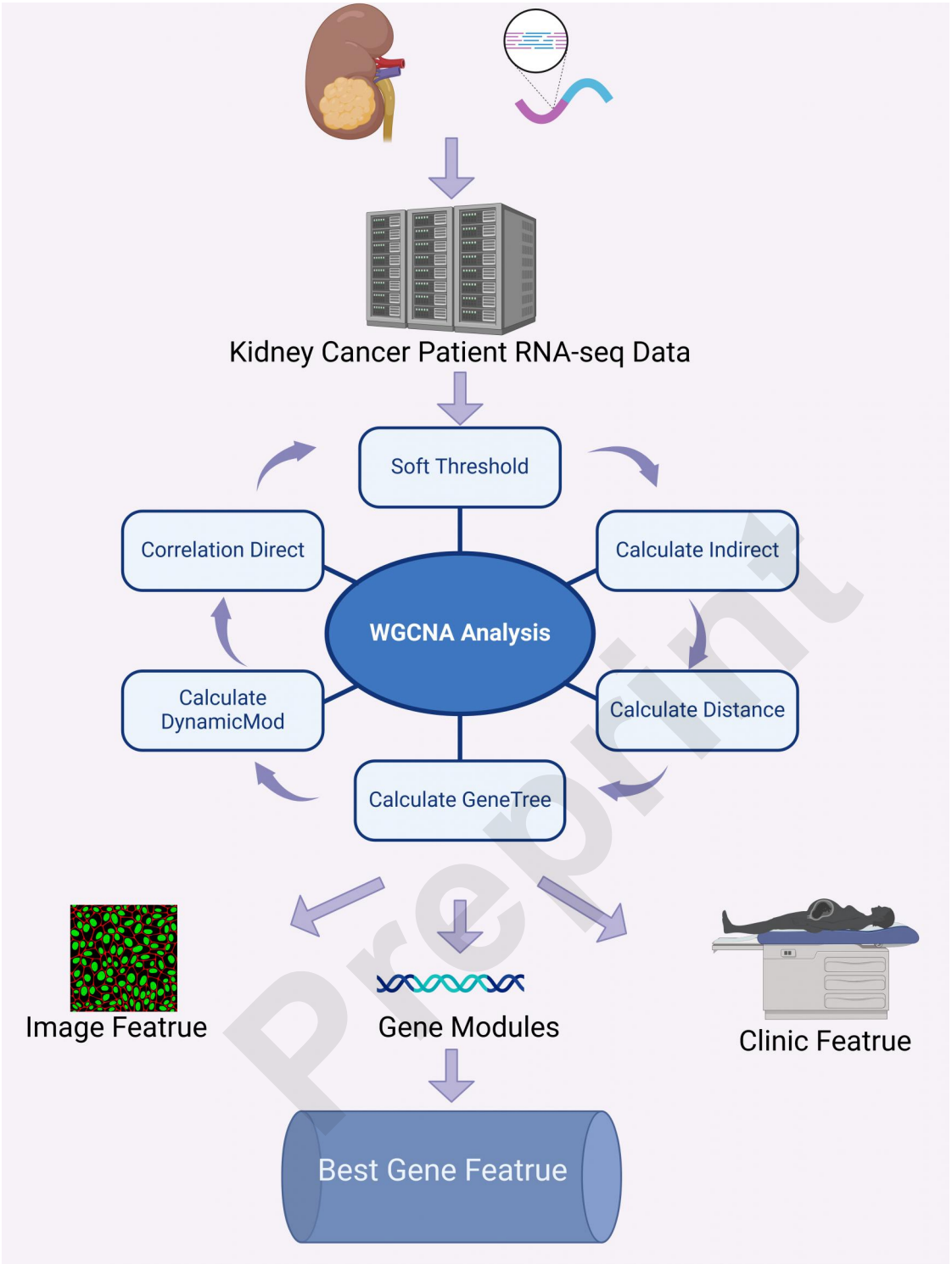



Figure 3. Workflow for Identifying Optimal Gene Features by Integrating Imaging and Clinical Features.

Note: The WGCNA primarily involves six steps: calculating the direct correlation between genes (gray module in the figure), determining the soft threshold for gene modules (gray module), calculating the indirect correlation between genes (purple module), calculating the Euclidean distance between gene modules (blue module), obtaining the gene tree pathway between gene blocks (orange module), and dividing and cutting gene modules (yellow module). These methods are combined to classify all genes, facilitating the subsequent integration of imaging data and clinical features for analysis.

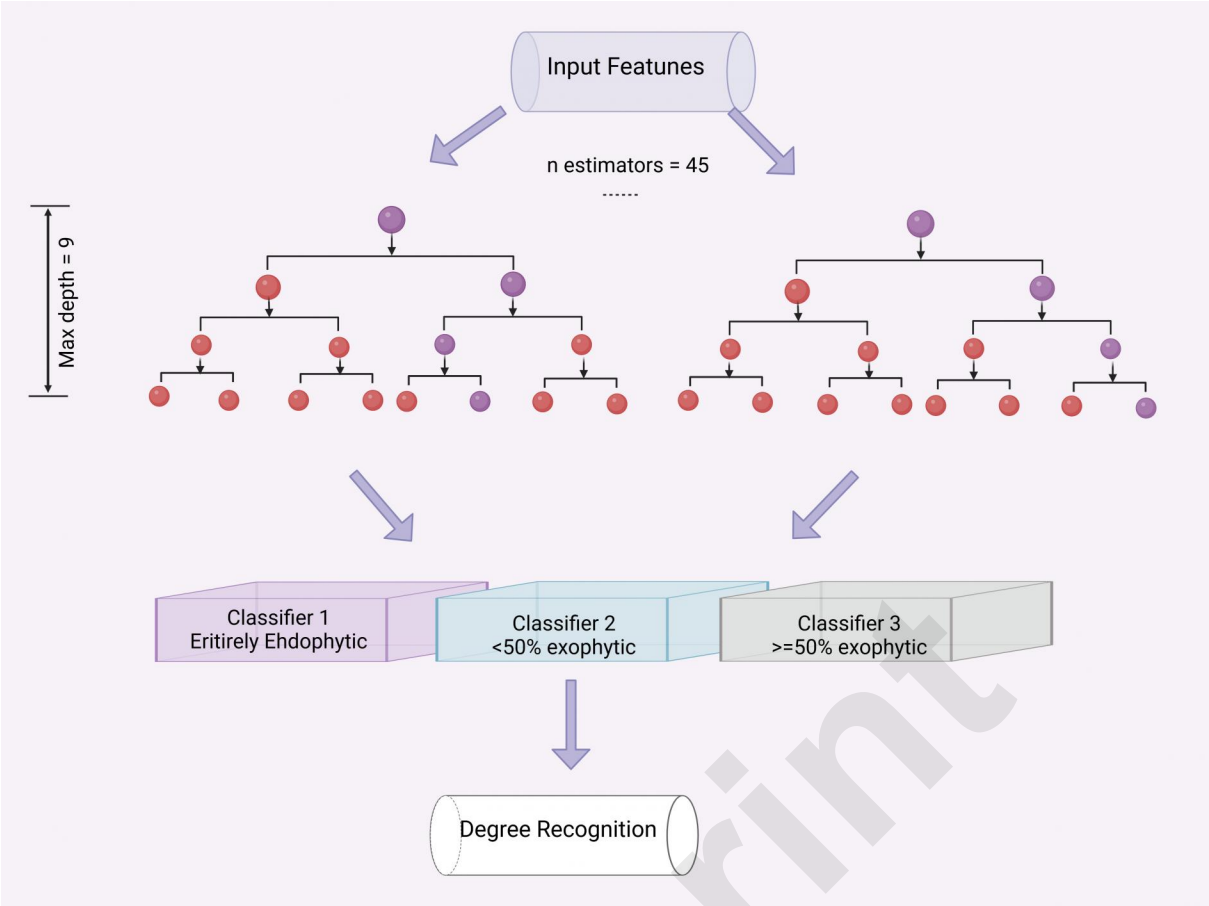



Figure 4. Prognostic Model Based on Random Forest Integrating ccRCC Genes, Imaging, and Clinical Features.

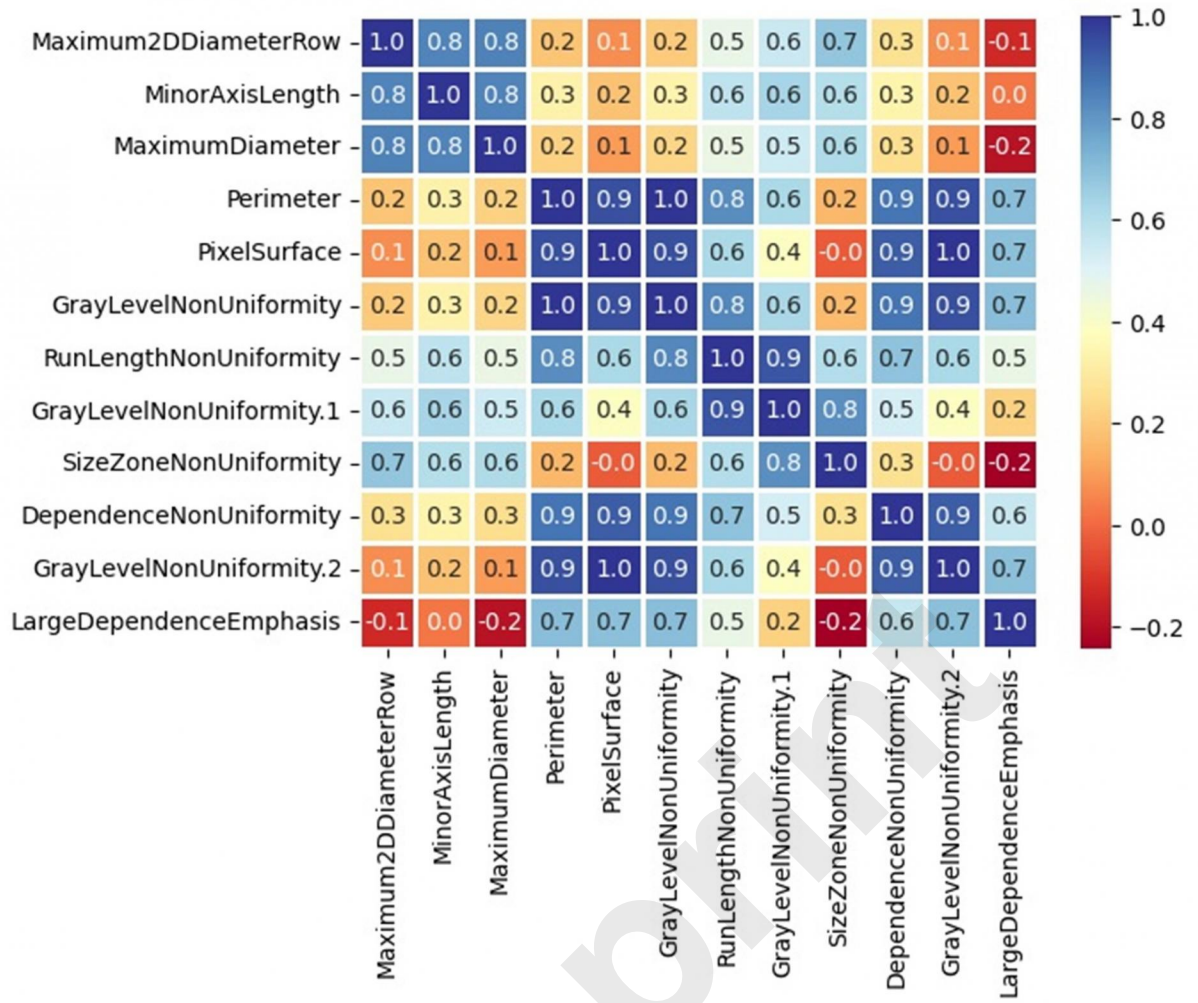



Figure 5. Similarity Between Features Selected Through Lasso Regression.
Note: The 12 selected features are as follows: Maximum 2D Diameter Row (maximum two-dimensional diameter of the tumor in the sagittal plane), Minor Axis Length (short axis length of the tumor), Maximum Diameter (maximum tumor diameter), Perimeter (tumor region perimeter), Pixel Surface (number of surface pixels in the tumor region), Gray-Level Non-Uniformity 1~3 (non-uniformity of gray level), Run Length Non-Uniformity (non-uniformity of running length), Size Zone Non-Uniformity (non-uniformity of area size), Dependence Non-Uniformity (dependent on non-uniformity), and Large Dependence Emphasis (gray-level dependence matrix). Each of these features is crucial for prognosis, with a significant impact factor.

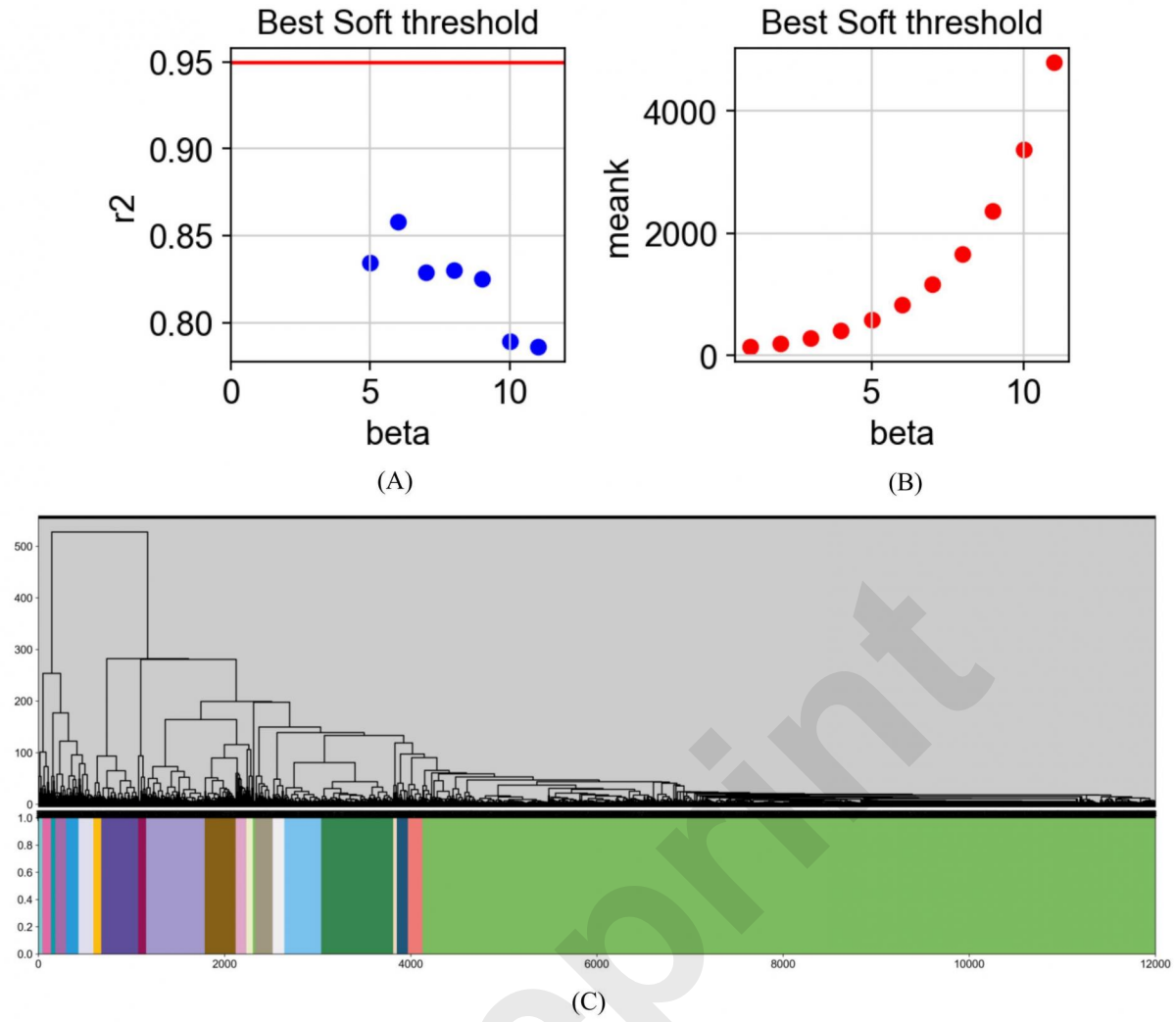





Figure 6. Visualization of Gene Analysis Based on WGCNA.

Note: The WGCNA method transforms the direct correlation matrix into an indirect correlation matrix to calculate the soft threshold. Using this soft threshold, the original correlation network is converted into a scale-free network. (A) and (B) present the scale-free network in different coordinate systems. (A) the vertical axis shows the evaluation metric r2 for the scale-free network. The closer r2 is to 1, the more the network approaches a scale-free structure, with r2 typically required to be greater than 0.8 or 0.9; (B) the vertical axis represents the mean connectivity, which decreases as the β value increases. By combining these two metrics, we typically select the β value where r2 first reaches 0.8, 0.9, or higher. This β value is then used to convert the correlation matrix into an adjacency matrix, which represents the connections between nodes in the network and is key to building the scale-free network; (C) shows a dynamic tree cut method applied to the topological overlap matrix based on the scale-free network, revealing a co-expression network analysis where each gene module contains co-expressed genes that share similar gene characteristics.

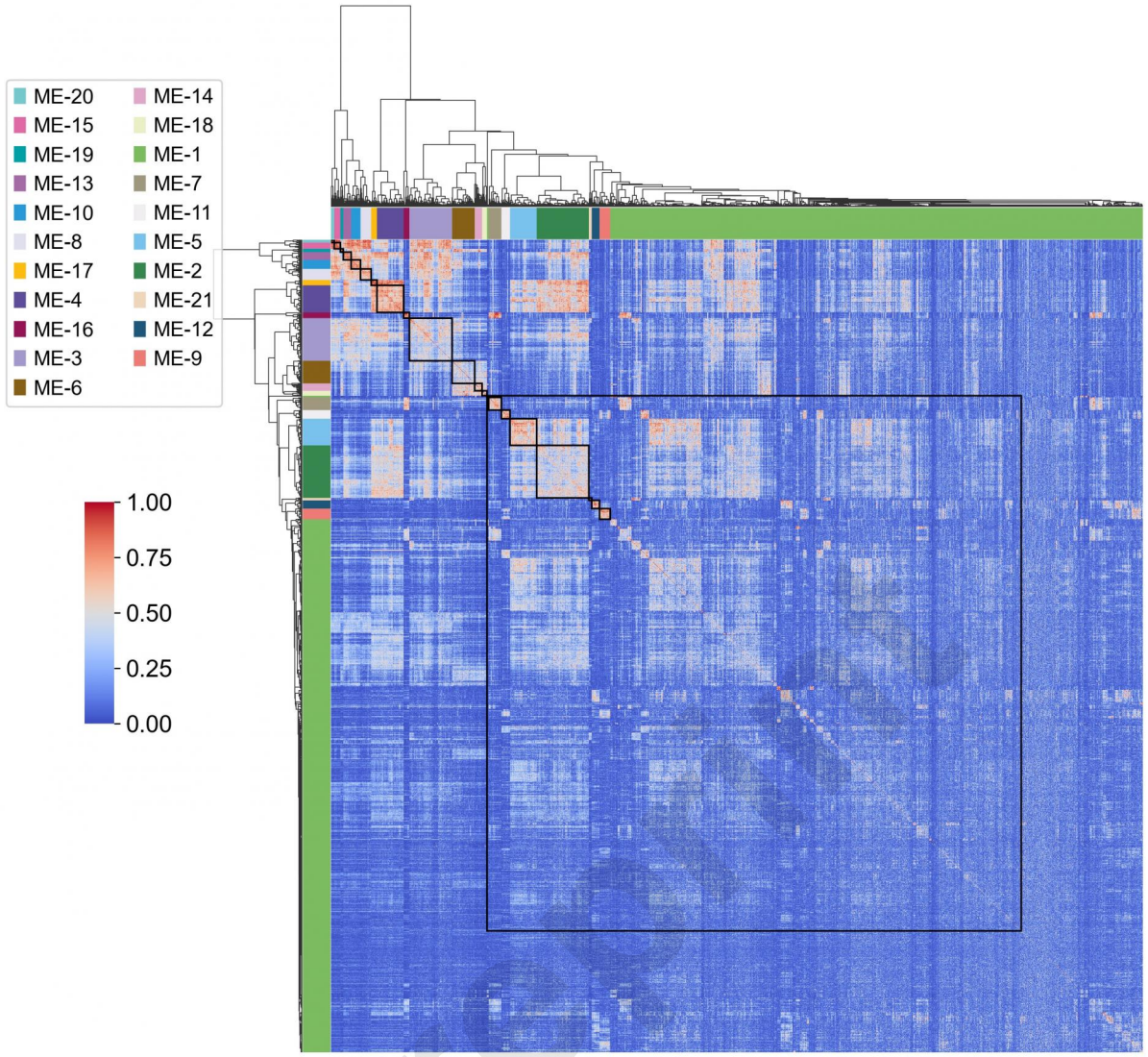



Figure 7. Visualization of the Topological Overlap Matrix Between Gene Modules.
Note: The values in the topological overlap matrix range from 0 to 1, with higher values indicating greater similarity between gene modules. It is evident that smaller gene modules exhibit higher similarity, while larger gene modules show lower similarity. This suggests that smaller gene modules may share similar functions or regulatory mechanisms, whereas larger gene modules might have distinct functions or regulatory mechanisms.

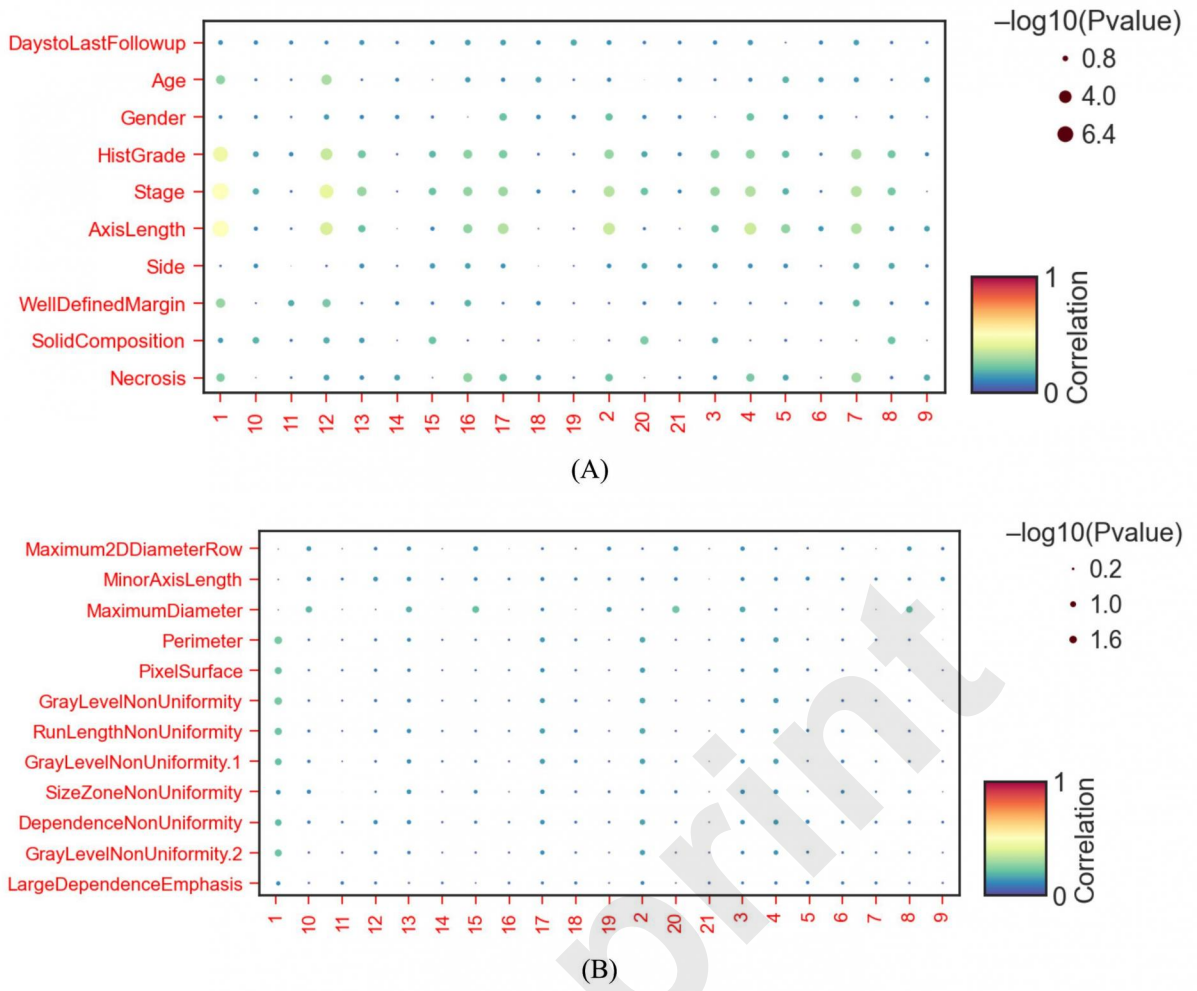



Figure 8. Evaluation of the Model.
Note: (A) shows the correlation analysis between gene modules and clinical features; (B) shows the correlation analysis between gene modules and imaging data. In both figures, "Correlation" represents similarity, with darker red indicating greater similarity. The size of the circles represents the expression level of the gene modules for the given trait, with larger circles indicating higher expression levels.

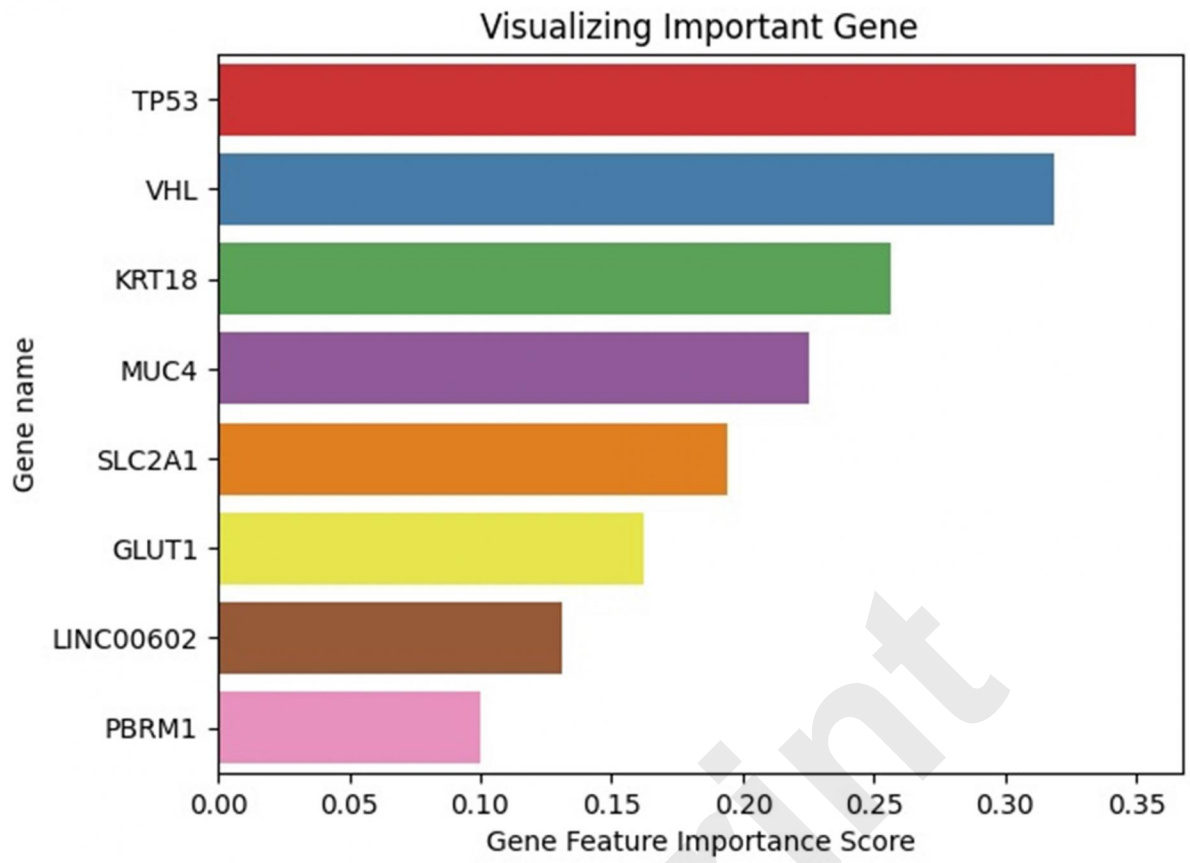



Figure 9. Visualization of the Confusion Matrix for Model Results.

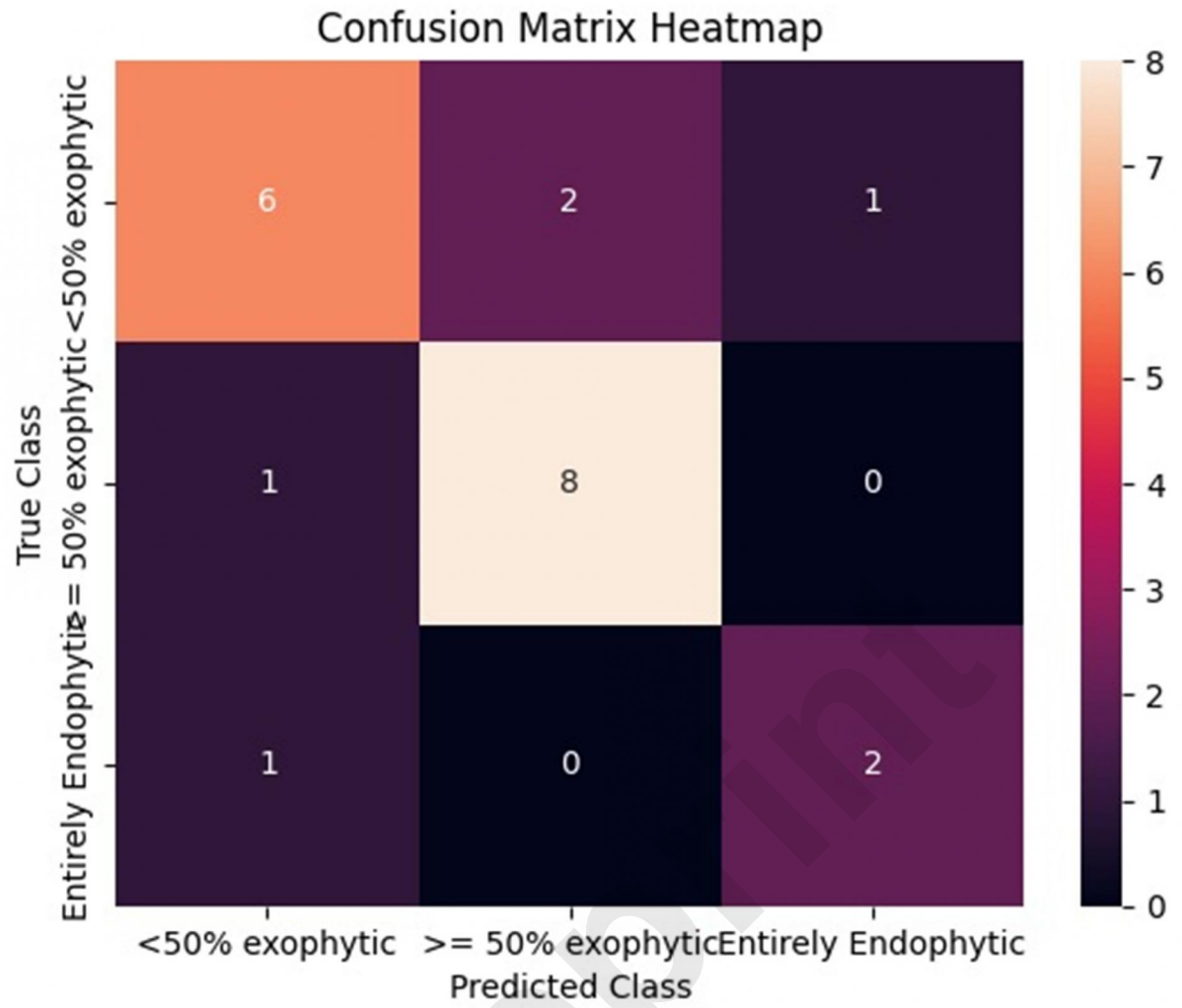



Figure 10. Contribution of the Top Eight Genes with the Highest Weight in the Prognostic Model.