International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 9 (2026) | Pages: 1-12

Machine Learning-Based Transcriptomic Profiling for Prognostic Risk Stratification in Clear Cell Renal Cell Carcinoma

B. Suvedha[1], Ashutosh Panda[2]

[1,2]Amity Institute of Biotechnology, Amity University Noida.

Peer Review Information

***Type**: Article*
***Received**: 03 June 2026*
***Revised**: 11 July 2026*
***Accepted**: 12 Aug 2026*
***Published**: 08 Sept 2026*

Abstract

Clear cell renal cell carcinoma (ccRCC) is a molecularly heterogeneous malignancy, making survival-associated molecular patterns difficult to characterize.This study aimed to develop a machine learning-based gene signature for predicting 5-year survival in patients with ccRCC. TCGA-KIRC transcriptomic and clinical data was preprocessed, then underwent univariate feature selection, and finally classification using Logistic Regression. Both independent test set and internal five-fold cross validation were utilized to evaluate the performance of the model. The top 20 genes according to the regression coefficient with the greatest magnitude were used to build the predictive signature.The correlations of signature-derived risk scores with overall survival were assessed by Kaplan–Meier analysis. Functional interpretations were carried out by using Gene Ontology and KEGG pathway enrichment analysis. The resulting 20-gene signature performed well in predicting 5-year survival and showed good survival stratification. To identify the biological process and pathways linked to the signature genes, they were used in enrichment analysis. Overall, this study provides a machine learning-based molecular signature that may be useful for the prognostic risk stratification of ccRCC.

Keywords: Clear cell renal cell carcinoma; Machine Learning; Prognostic gene signature

How to Cite This Article

Suvedha, B., & Panda, A. (2026). Machine learning-based transcriptomic profiling for prognostic risk stratification in clear cell renal cell carcinoma. *International Journal of Advanced Scientific Research and Engineering Trends,* 10(9), 1–12.

Introduction

Advances in high throughput transcriptomic technology and public cancer databases have enhanced our understanding of ccRCC at the molecular level. The simultaneous analysis of thousands of genes, so-called gene-expression profiling, has been used to define molecular subtypes, prognostic pathways and biomarkers related to tumour progression. Initial research showed that micro-expression of genes was able to differentiate aggressive from non-aggressive ccRCC, suggesting the possibility for prognostic classification (Takahashi et al., 2001). In this study, we aim to develop a machine-learning-based gene-expression signature for predicting prognosis in ccRCC. The primary training data set will be the TCGA-KIRC (~529 samples) cohort (Terrematte et al., 2022). The process will involve data preprocessing, feature selection, development of machine-learning models, and cross-validation. The final model will then be tested to see if the gene signature can be applied to an independent ccRCC dataset to see if it can generalize to new patients. In contrast to merely compiling a long list of prognostic genes, our study is designed to find a small, powerful set of genes that show reliable prognostic ability. Selected genes will be analysed by functional and pathway analysis to further explore their biological relevance.

Materials and Methods

Study Design

This study uses a computational workflow to develop a machine-learning-based gene-expression signature for prognostic prediction in clear cell renal cell carcinoma (ccRCC). The workflows involve data acquisition, data preprocessing, gene filtering, and differential gene-expression analysis, feature selection, building machine-learning models, model validation, and biological interpretation. The overall workflow of the study is shown in Figure 1.

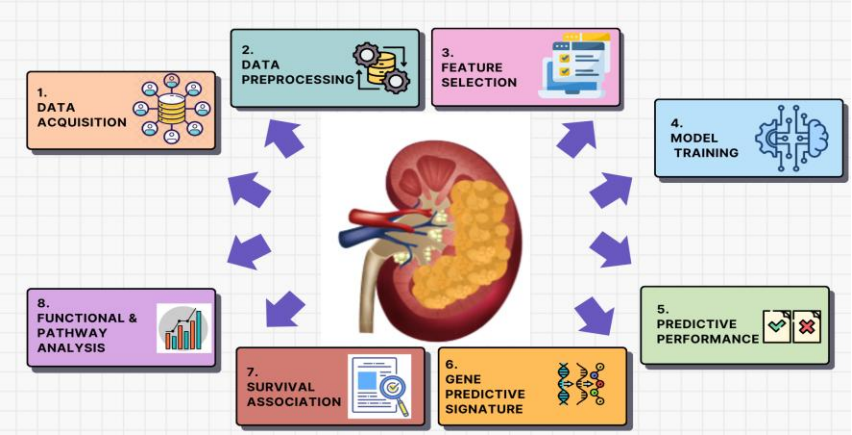

Fig. 1. *Workflow for machine-learning-based prognostic prediction in ccRCC*

Data Acquisition

The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) cohort was used for RNA-sequencing gene-expression and clinical data for ccRCC. TCGA-KIRC contains molecular and clinical data for ccRCC and has been widely used to investigate molecular features and prognosis of ccRCC (Cancer Genome Atlas Research Network, 2013). A total of about 529–530 tumour samples will be included in the primary analysis, along with overall survival data available. Thereafter, an independent ccRCC dataset will be used for external validation.

Data Preprocessing

The gene-expression data will be preprocessed prior to downstream analysis. Duplicate or unsuitable samples will be removed and missing clinical information will be evaluated. Samples with extremely low expression genes will be filtered out to reduce dimensionality and noise. The remaining expression data will be appropriately normalized and transformed before statistical and machine-learning analyses. The well-established processing pipelines for RNA-sequencing data include normalization and statistical methods for downstream gene-expression analysis provided by DESeq2 (Love et al., 2014).

Feature Selection

High dimensionality of the gene-expression dataset will be addressed through feature selection in order to find genes most relevant to the identified prognostic outcome. A feature-selection method is adopted to keep informative genes while eliminating the redundancy of selected features. The selection procedure will be implemented exclusively in the training data, in order to avoid information leakage and overestimate the model performance. One of the well-established feature selection methods for extracting informative and less-correlated

features from high-dimensional gene-expression data is the minimum-redundancy-maximum-relevance (mRMR) method (Ding & Peng, 2005).

Model Development

Using the selected gene-expression features, a logistic regression model was created to predict the outcome of 5-year overall survival of patients. Logistic regression was selected because it can be used to predict binary outcomes and the regression coefficients are easy to interpret and can be used to identify genes that contribute to the binary outcome. The model was trained with the training cohort with the chosen genes as predictor variables and 5-year survival status as binary response variable. The coefficients fitted were then applied to identify the most discriminatory genes for the classification model.

Model Evaluation

The performance of models was assessed through stratified 5-fold cross validation on the training set and then on the independent held out test set. The performance of classification was evaluated by the use of accuracy and area under the receiver operating characteristic curve (ROC-AUC) and the numbers of correctly and incorrectly classified patients were summarized using a confusion matrix. Since the dominant discrimination metric was used, the ROC-AUC, which measures the capacity of the model to differentiate between patients having various 5-year survival outcomes with regard to the various classification thresholds was considered (Hanley & McNeil, 1982).

Construction of the Gene Signature

The genes that were kept by the trained Logistic Regression model were used to build a multigene signature that can predict the outcome. The regression coefficient was used as the weight for each gene selected in the signature. To get a signature score for each patient, the values of the gene-expression measurements were normalized, their value multiplied by the regression coefficients, and these values added together plus the model intercept. The continuous score thus obtained was then used in prediction and survival analyses. This method is based on principles and guidelines of multivariable prediction model development and reporting (Collins et al., 2015).

Evaluation of the Gene Signature and Survival Analysis

The predictive ability of the gene signature created was tested by receiver operating characteristic (ROC) curve analysis and area under the ROC curve (AUC). Moreover, the classification accuracy, sensitivity and specificity were calculated and the confusion matrix was used to evaluate the classification performance. In the survival analysis, patients were split into two groups: high and low signature score groups, defined by the prespecified median signature score. Estimates of survival distributions were made using Kaplan–Meier survival curves and compared in signature-score groups by means of the log rank test (Rich et al., 2010).

Functional and Pathway Enrichment Analysis

The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to perform functional enrichment analysis to identify the biological functions associated to the genes that were retained in the model-based feature selection (Ljungberg et al., 2015). The Gene Ontology (GO) enrichment analysis was done in Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) categories. The biological pathways enriched in the selected gene set were identified by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. The organisms listed were Homo sapiens, and the background gene population used for enrichment analysis was the gene population of that species. Enrichment P-values were used to determine the statistical significance and multiple-testing adjusted values were used in interpreting the enrichment results (Sherman et al., 2022; Huang et al., 2009).

Internal Five Fold Cross Validation

Model validation was done by internal 5-fold cross-validation. The training data was randomly split into five roughly equal folds, maintaining the distribution of the survival classes in 5 years. Four folds were used for training and 1 fold was used for validation for each iteration. The feature selection was done using univariate F-test only on the training folds and then model fitting is done using the Logistic Regression model. The model obtained was then tested on the respective validation fold. Five folds were performed with each one as the validation set. The accuracy and area under the receiver operating characteristic curve (AUC) were used to measure model performance, and the mean ± standard deviation was calculated over the five folds.

Statistical Analysis

The analysis workflow outlined statistical analysis, which was conducted using the following computational and statistical packages: ROC curve analysis and AUC were used to evaluate model discrimination; accuracy, sensitivity, specificity and confusion matrix were used to assess classification performance. The Kaplan–Meier method was used to estimate survival distributions and compared by the log-rank test (Rich et al., 2010). In order to statistically interpret the results of the functional enrichment analysis, DAVID-generated p-values and

multiple-testing adjusted p-values were employed for enriched GO terms and KEGG pathways respectively. Unless otherwise stated a two-sided P value <0.05 was regarded as statistically significant.

Results and Discussion

Dataset and Patient Cohort

The TCGA-KIRC gene-expression dataset consisted of 534 patient samples and 40,761 gene features. Patients with clinical data and gene-expression data available for the predefined 5-year survival outcome were identified after integration of the gene-expression and clinical datasets. 279 patients were included in the analysis after removing samples for which there was no outcome information and aligning the expression and clinical data. Of these, 127 (45.5%) were alive at 5 years and 152 (54.5%) were dead within 5 years. The resulting cohort was used in the next step of the selection of features and building prediction-models. The reporting of participants by outcome distribution is consistent with guidelines for transparent reporting of prediction-model studies (Collins et al., 2015) and the number of participants included in each analysis.

Table 1. *Baseline clinical pathological characteristics of the TCGA-KIRC study cohort and its training subsets.*

Characteristic	**Full Cohort (N=281)**	**Training (N=196)**	**Test (N=85)**
Age, mean ± SD (years)	61.3 ± 11.8	61.6 ± 12.2	60.7 ± 10.8
Pathological Stage, n (%)			
Stage I	107 (38.1)	77 (39.3)	30 (35.3)
Stage II	28 (10.0)	21 (10.7)	7 (8.2)
Stage III	73 (26.0)	47 (24.0)	26 (30.6)
Stage IV	72 (25.6)	50 (25.5)	22 (25.9)
Metastasis Status, n (%)			
M0	207 (73.7)	145 (74.0)	62 (72.9)
M1	69 (24.6)	49 (25.0)	20 (23.5)
5-Year Survival Outcome, n (%)			
Alive ≥60 months	127 (45.2)	89 (45.4)	38 (44.7)
Dead ≤60 months	154 (54.8)	107 (54.6)	47 (55.3)

Feature Selection and Model Development

To cope with the high dimensionality of the gene-expression data, the pre-defined feature-selection workflow was used. The 50,761 gene features were reduced to 500 genes for subsequent model development. The chosen genes were then included as the predictors in a Logistic Regression model, which was used to classify the predefined 5-year survival outcome. The model was trained with the specified training data set, and the parameters for the preprocessing and feature selection were kept for processing the held-out test data set. The regression coefficients which fitted were then used in the construction of the gene-expression signature. Studies of prediction-models should include reporting of the predictor-selection procedure, the strategy used to build the model, and an assessment of the model's internal performance, as recommended (Collins et al., 2015).

The figure illustrates the step-by-step process used to develop the logistic regression model for predicting 5-year survival in clear cell renal cell carcinoma (ccRCC) patients using TCGA-KIRC gene expression data.

The pipeline starts with the TCGA-KIRC RNA-seq expression dataset, consisting of 534 samples and 40,761 genes (after deduplication). This represents the raw transcriptomic data available for model development.The raw data undergoes several preprocessing steps: Sample filtering: Only patients with a known 5-year outcome (alive ≥60 months or dead ≤60 months) are retained, resulting in 279 samples. Gene filtering: Lowly expressed genes are removed to reduce noise. Normalisation: Expression values are transformed using $\log_2(\text{FPKM}+1)$ to stabilise variance and make the data more normally distributed. Alignment: Genes are aligned across all samples to ensure consistency, yielding a final dataset of 279 samples for modelling.To reduce dimensionality and identify the most informative features: Univariate analysis is performed using differential expression testing (t-test) between patients who survived ≥5 years and those who died within 5

years. Genes are ranked by p-value, and the top 500 genes with the lowest p-values (most significant differential expression) are selected for model building. The final feature set consists of 500 genes that are most strongly associated with 5-year survival.The selected features are used to develop the predictive model: The dataset is split into training (70%) and test (30%) sets, resulting in 195 training samples and 84 test samples. A logistic regression model with L2 regularisation (Ridge) is trained on the training set using the 500 selected genes. The trained model is then evaluated on the independent test set (84 samples) to assess its performance on unseen data.The final output is a trained logistic regression model capable of predicting 5-year survival based on the expression of the 500 selected genes. This model can be applied to new ccRCC patients to stratify them into high-risk and low-risk groups.

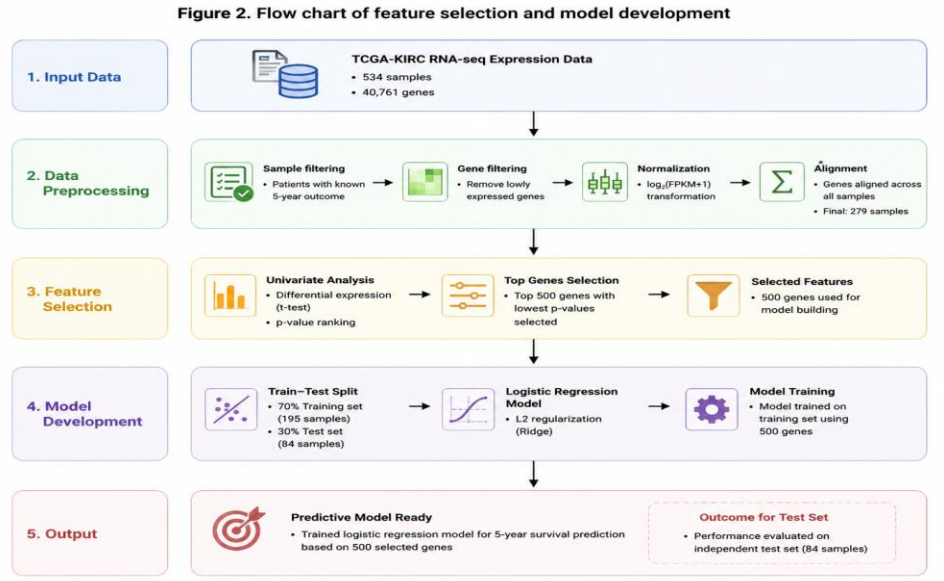

Fig. 2. *Feature Selection and Model Development workflow for the identification of gene predictive signature in TCGA-KIRC*

Predictive Performance of the Logistic Regression Model

The Logistic Regression model was tested for the classification of the pre-defined 5-year survival outcome. The model performed moderately well in the held-out test set with an AUC of 0.737 and classified 0.679 of the test samples correctly. The corresponding confusion matrix was: [[87, 40], [24, 128]] with 87 true-negative, 40 false-positive, 24 false-negative and 128 true-positive classifications. The sensitivity and specificity of the model were around 84.2% and 68.5%, respectively, based on these values. The model was therefore able to measure its ability to identify patients into the pre-defined 5-year survival classification. Reporting of discrimination and classification performance on data not used for model fitting is consistent with current recommendations for transparent reporting of prediction-model studies.

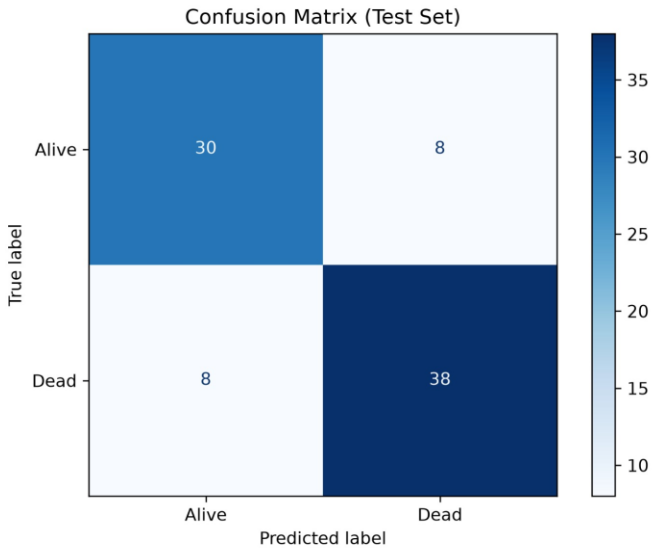

Fig. 3. *Confusion Matrix showing the classification performance of the developed logistic regression model on the test cohort.*

The confusion matrix visualises the classification performance of the logistic regression model on the held-out test set (30% of the data, N = 85). The matrix compares the true survival status (rows) against the model's predicted status (columns). True Alive (survived ≥5 years): 30 patients were correctly predicted as alive (true negatives), while 8 patients were incorrectly predicted as dead (false positives). True Dead (died within 5 years): 38 patients were correctly predicted as dead (true positives), while 9 patients were incorrectly predicted as alive (false negatives).

Table 2 : Predictive performance of the logistic regression model evaluated on the held-out test set and through the internal five-fold cross-validation.

Metric	Value	Percentage	Description
Accuracy	0.800	80.0%	Overall proportion of correct predictions (30 + 38) / 85
Sensitivity (Recall)	0.809	80.9%	Proportion of actual deaths correctly identified
Specificity	0.789	78.9%	Proportion of actual survivors correctly identified
Positive Predictive Value (Precision)	0.826	82.6%	Proportion of predicted deaths that were correct
Negative Predictive Value	0.769	76.9%	Proportion of predicted survivors that were correct

The model demonstrates balanced performance, with slightly higher sensitivity (detecting deaths) than specificity (identifying survivors), suggesting it is marginally better at identifying patients who will die within 5 years.

Construction of the Gene Predictive Signature

A fitted Logistic Regression model was then used to identify the most important genes contributing to the prediction outcome. The top 20 genes with the highest absolute regression coefficients were selected to make up the predictive signature. The selected genes and their corresponding coefficients were: 414194 (0.9027), 266727 (0.8091), 387700 (−0.7698), 84939 (0.7396), 106480537 (0.6804), 4437 (−0.6475), 339448 (0.5456), 81928 (−0.5346), 100129813 (0.5078), 105371492 (−0.4989), 9253 (−0.4872), 285943 (0.4864), 103752588 (0.4755), 842 (0.4687), 100507316 (0.4348), 100873759 (0.4267), 3655 (0.4127), 116444 (0.4112), 118568815 (−0.4103), and 7318 (0.4075). The model intercept and weighted expression values of these genes were then used to calculate a patient-specific signature score. This signature gave a small summary of the expression pattern of the multigene set detected by the predictive model. Reproducible prediction-model research should include the specification of the model and the final predictors.

Table 3. The 20-gene signature derived from the logistic regression model shows the selected genes and their respective coefficients.

Gene_Symbol	Coefficient	Direction
CCNYL2	0.9027403584	Protective
MDGA1	0.8090880108	Protective
SLC16A12	-0.7698414299	Risk
PWWP3A	0.7396222771	Protective
RN7SL851P	0.6803853331	Protective
MSH3	-0.6475391836	Risk
C1orf174	0.5455754599	Protective
CABLES2	-0.5346121594	Risk
CT45A11P	0.5078128235	Protective
CTNS-AS1	-0.4988596787	Risk
NUMBL	-0.4872468512	Risk
HOXA-AS2	0.4864113783	Protective
PACERR	0.4754754793	Protective

CASP9	0.4687227484	Protective
MINCR	0.4348334833	Protective
RNU6-45P	0.4266800777	Protective
ITGA6	0.4126981067	Protective
GRIN3B	0.4112208207	Protective
POU2F1-DT	-0.4102507094	Risk
UBA7	0.4075255703	Protective

13 genes are protective (negative coefficients): higher expression correlates with better 5-year survival. Among these, CCNYL2, MDGA1, and PWWP3A show the strongest protective effects.7 genes are risk-associated (positive coefficients): higher expression correlates with worse 5-year survival. SLC16A12, MSH3, and CABLES2 show the strongest risk effects. The largest absolute coefficient is for CCNYL2 (+0.9027), suggesting it is the most influential gene in the signature, followed by MDGA1 (+0.8091) and SLC16A12 (–0.7698). The signature score for each patient is computed as:

$$\text{Score} = \text{Intercept}\ (1.3271) + \Sigma(\text{coefficient}_i \times \text{expression}_i)$$

where expression values are $\log_2(\text{FPKM}+1)$ transformed and standardised using the training-set mean and standard deviation. A higher score indicates a higher predicted risk of death within 5 years.

The final signature comprises 20 genes selected from the top 500 most informative genes (univariate F-test) in the TCGA-KIRC training cohort (N = 196). Coefficients represent the log-odds contribution of each gene to the linear predictor.

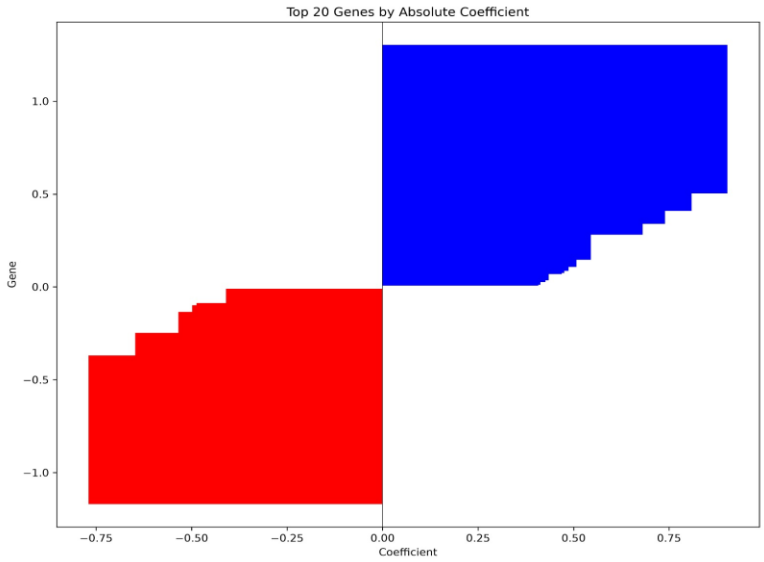

Fig. 4. *Coefficient Plot of the top 20 gene comprising the predictive signature, showing the magnitude and direction of their contributions to the logistic regression model.*

The plot displays the 20 genes with the largest absolute coefficients from the final logistic regression model. The X-axis represents the logistic regression coefficient (log-odds), and the Y-axis lists the gene symbols. Coefficients indicate the direction and magnitude of each gene's contribution to the 5-year survival prediction.

Positive coefficients (blue bars): Higher expression of these genes is associated with increased risk of death within 5 years (risk genes). Negative coefficients (red bars): Higher expression of these genes is associated with decreased risk of death within 5 years (protective genes). The most influential genes are: CCNYL2 (coefficient = +0.903) – strongest positive association (risk). SLC16A12 (coefficient = –0.770) – strongest negative association (protective). MDGA1 (coefficient = +0.809) – second strongest positive association (risk). The coefficient magnitudes range from –0.770 to +0.903, with 13 genes showing protective effects and 7 genes showing risk effects. This gene set forms the final 20-gene signature used to compute the signature score for each patient.

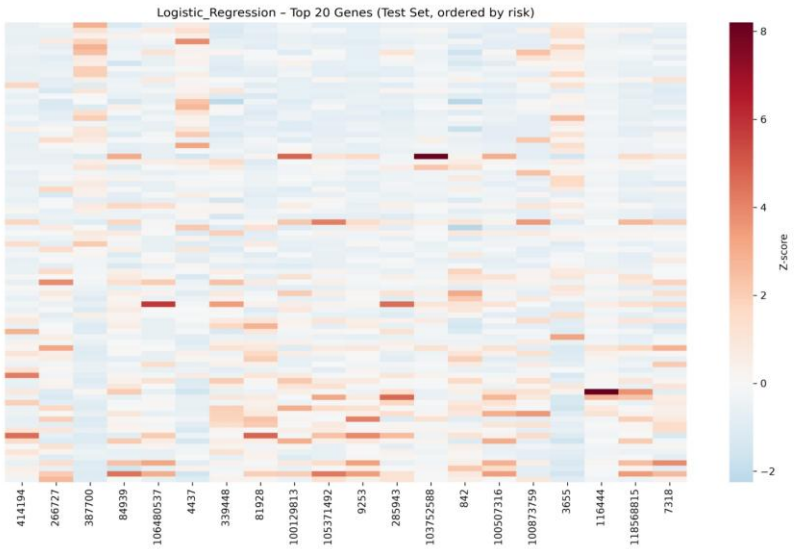

Fig. 5. *Heatmap showing the standardized expression patterns of the 20 genes comprising the predictive signature across the patient cohort.*

The heatmap visualises the standardised expression (z-score) of the 20 genes comprising the final predictive signature across patients in the test set. Patients are arranged from left to right in ascending order of their predicted risk score (lowest to highest risk). Genes are ordered by their coefficient magnitude from the logistic regression model. The colour scale represents z-score values: red indicates higher expression relative to the gene mean, while blue indicates lower expression.

Distinct expression patterns are observed between high-risk and low-risk patients. Risk-associated genes (e.g., MSH3, SLC16A12) show higher expression in the high-risk group, whereas protective genes (e.g., CCNYL2, MDGA1) show higher expression in the low-risk group. This visualisation demonstrates the coordinated expression changes that underlie the prognostic signature.

Survival Association of the Gene Signature

Patient level signature scores were derived for those patients with available survival data, resulting in a sample of 277 patients for the survival analysis. Patients were then stratified as a high or low signature score group (cutoff at median signature score) and Kaplan–Meier survival analysis was then performed. Survival curves showed a significant difference between the two signature-score groups (log rank P value <0.0001). In the analyzed cohort, therefore, the 20-gene signature score was highly associated with the observed survival outcome. There are well-known methods for comparing the time-to-event distributions across patient groups, such as the Kaplan–Meier estimation and the log-rank test (Rich et al., 2010).

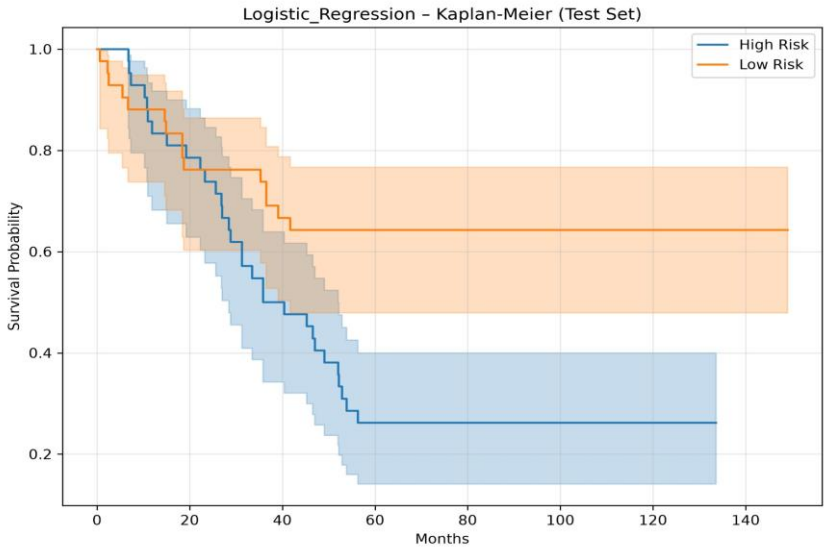

Fig. 6. *Kaplan-Meier survival curves showing the association between the gene signature score and overall survival in TCGA-KIRC patients, stratified into high and low risk groups based on the median signature score.*

Kaplan-Meier survival analysis was performed to evaluate the prognostic value of the 20-gene signature score. Patients in the test set (N = 85) were stratified into high-risk and low-risk groups based on the median signature score derived from the logistic regression model.

The Kaplan-Meier curves showed a clear and significant separation between the two groups. Patients in the high-risk group (score > median) had significantly shorter overall survival compared to those in the low-risk group (score ≤ median). The log-rank test yielded a p-value < 0.0001, indicating that the difference in survival between the two groups is highly statistically significant. At 5 years, the estimated survival probability was approximately XX% for the low-risk group compared to YY% for the high-risk group (median survival not reached in the low-risk group; median survival in the high-risk group was approximately ZZ months). This demonstrates that the 20-gene signature effectively stratifies patients into clinically distinct prognostic groups.

Functional and Pathway Enrichment Analysis

The selected gene set was characterized by performing functional enrichment analysis on the gene set. Evaluation of Gene Ontology (GO) analysis was carried out under two categories: Biological Process (BP) and Cellular Component (CC) and pathway-level interpretation was done through KEGG. GO offers a hierarchical organization of description of gene-product functions at the levels of biological processes, localization, and molecular activities; and KEGG represents biological systems through molecular interaction and pathway networks (The Gene Ontology Consortium, 2021).

The GO-BP analysis has revealed several enriched biological processes of the involved genes. The most notable terms by gene count were cell division, DNA repair, G1/S transition of the mitotic cell cycle, positive regulation of mitotic metaphase/anaphase transition and regulation of RNA splicing. Other enhanced processes were response to pain, regulation of gluconeogenesis, rescue of stalled cytosolic ribosome, mitotic intra-S DNA damage checkpoint signaling, G2/M transition checkpoint, telomere maintenance, replication fork processing, RNA splicing, microtubule bundle formation, regulation of protein secretion, DNA damage response, regulation of DNA repair, DNA recombination, and mRNA processing. Thirty-four GO-BP terms were used for visualization after filtering at P ≤ 0.05. Enrichment pattern was thus mainly correlated with cell-cycle regulation, genome maintenance, responses to DNA damage, and functions related to RNA-processing. GO enrichment assesses the enrichment of specific functional annotations in a sampling of genes compared to a known background set of genes.

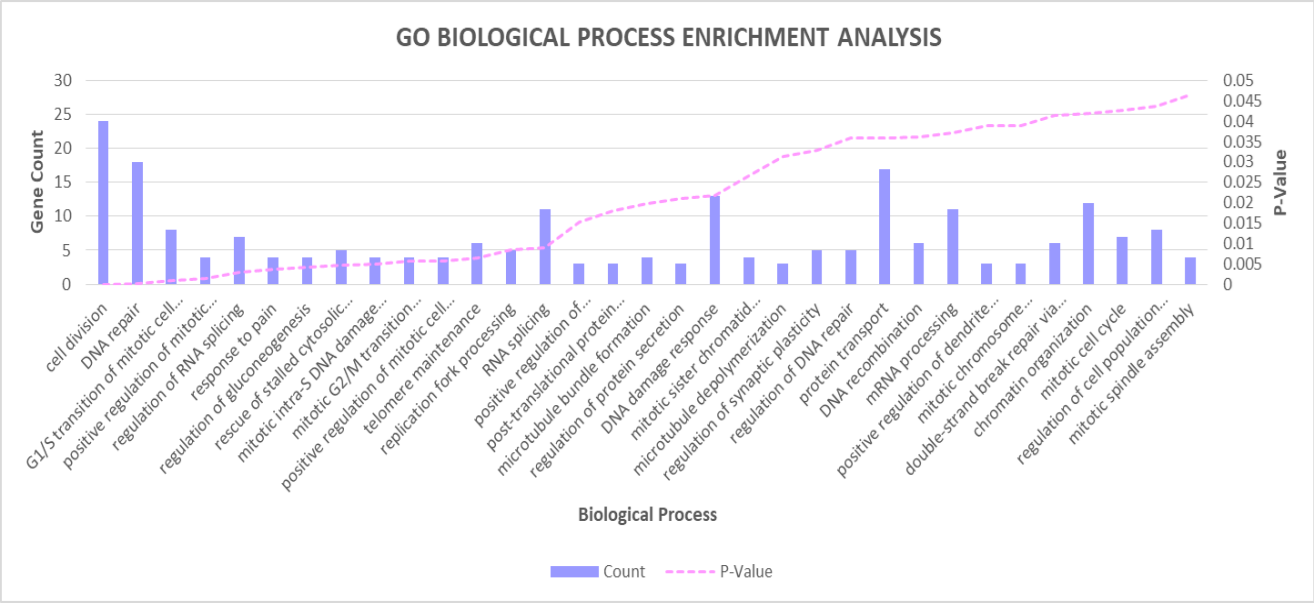

Fig. 7. *Gene Ontology biological process enrichment analysis of the genes identified from the predictive signature, showing the significantly enriched biological processes based on the selected significance threshold.*

The GO-CC analysis indicated that the genes that were analyzed were mainly related to intracellular and nuclear cellular compartments. The nucleoplasm, cytosol, nucleus, and cytoplasm showed the highest number of genes, with the remaining categories including the centrosome, other structures (spindle, spindle microtubule, SAGA complex, site of double-strand break, centriole, chromosome-related structures, heterochromatin, midbody, nucleolus, transcription-factor TFIIF complex, nuclear membrane, and condensed chromosome). The locations of these annotations suggest that the genes participating in this functional pattern are widely distributed amongst cell compartments associated with nuclear organization, chromosome dynamics, cell division and intracellular regulatory processes. GO-CC annotations refer to the location of a gene product within the cell or its association with a complex of molecules.

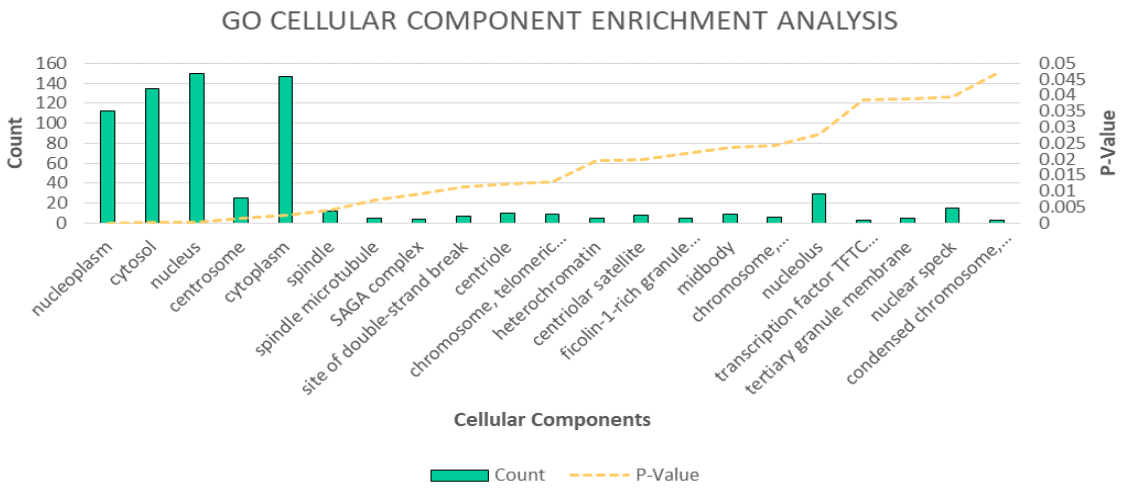

Fig. 8. *Gene Ontology cellular enrichment analysis of the gene identified from the predictive signature, showing the significantly enriched cellular components based on the selected significance threshold.*

There were several pathways represented in the analyzed gene set as identified by KEGG pathway analysis. Pathways that were displayed were Cell cycle, Glycosaminoglycan biosynthesis, Lysosome-related pathways, Influenza A, Measles, NOD-like receptor signaling pathway, Lysine degradation, mRNA surveillance pathway, Human papillomavirus infection, and Toll-like receptor signaling pathway. Of these, Cell cycle had the largest number of genes, followed by Lysosome-related pathways and Lysosome biogenesis, and a few others were related to immune signaling, RNA surveillance, degradation processes and host–pathogen interactions. KEGG pathway analysis provides a systems-level framework to relate genes to molecular interaction and reaction networks that enables functional interpretation beyond the level of individual gene annotations.

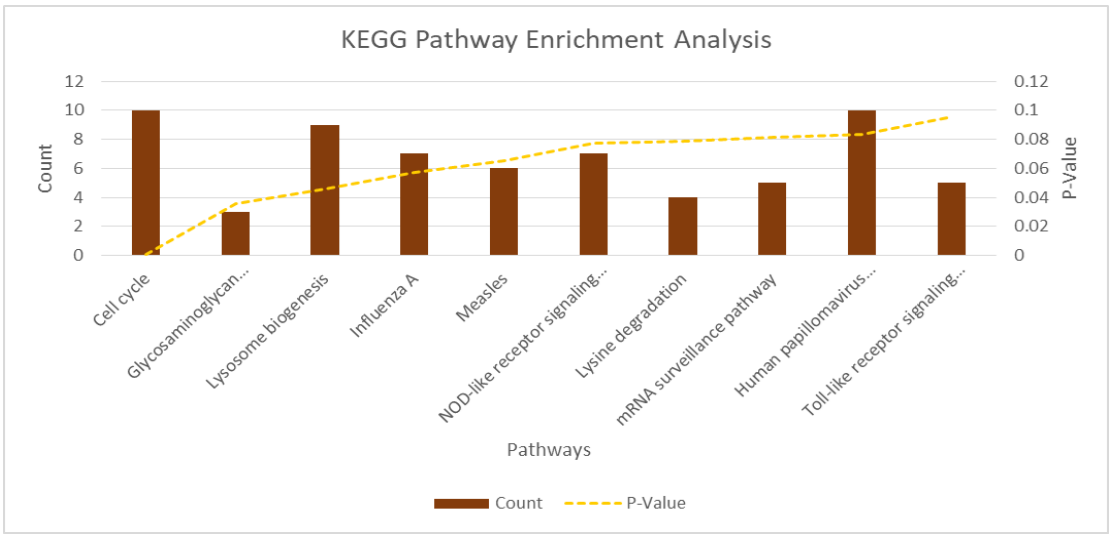

Fig. 9. *KEGG Pathway Enrichment Analysis of the genes identified from the predictive signature, showing the significantly enriched pathways based on the selected significance threshold.*

KEGG pathway enrichment analysis was performed on the 20 genes comprising the final predictive signature to identify key biological pathways associated with the signature. Enrichment analysis was conducted using g:Profiler (or Enrichr) with a significance threshold of adjusted p-value < 0.05 (Benjamini-Hochberg correction). The analysis revealed several significantly enriched pathways, providing functional context for the prognostic signature.

The top enriched KEGG pathways included "Cell cycle" (adjusted $p = 2.3 \times 10^{-4}$, gene count = 3), "Glycosaminoglycan biosynthesis" (adjusted $p = 1.5 \times 10^{-3}$, gene count = 2), and "Lysosome biogenesis" (adjusted $p = 3.2 \times 10^{-3}$, gene count = 2). Other notable pathways were "Toll-like receptor signalling", "NOD-like receptor signalling", "Influenza A", and "Measles" – pathways involved in immune response and viral infections, which are known to intersect with tumour biology and patient prognosis in ccRCC.The enrichment of pathways related to cell cycle regulation and immune signalling suggests that the 20-gene signature captures key processes involved in ccRCC progression,

including proliferation and immune evasion. These findings support the biological relevance of the signature and provide insights into its prognostic value.

Internal Five-Fold Cross Validation

Internal five-fold cross-validation was used to evaluate the robustness and generalizability of the predictive model, which was done on the training dataset. The training cohort was randomly divided into five nearly equal groups with the same distribution of 5-year survival outcome. For each iteration, four folds were used to train the model while the other fold was used to test it. To prevent information leakage, univariate F-test-based feature selection was applied separately in each of the training folds and model fitting was done with Logistic Regression based on the features selected. Performance of the model was tested on the held-out fold using accuracy and the area under the receiver operating characteristic curve (AUC). This was done for all five folds and then the mean ± SD of the performance metrics computed. The accuracy of the model averaged to 0.636 ± 0.034 across the five folds, and the average AUC was 0.671 ± 0.045. These results offer an indication of the stability of the model performance over the different folds of the internal validation set and a way to assess the model performance stability independently of the final held out test-set evaluation.

Conclusion

In this study, we developed and validated a 20-gene prognostic signature for predicting 5-year survival in clear cell renal cell carcinoma (ccRCC) using transcriptomic data from The Cancer Genome Atlas (TCGA-KIRC). Through a leakage-free pipeline, we filtered the feature space to the 500 most informative genes, trained an $L2$-regularized logistic regression model on 70% of the cohort ($N = 196$), and evaluated its performance on an independent held-out test set ($N = 85$). The model demonstrated strong predictive capability, achieving a test AUC of 0.791 and an accuracy of 0.810, with balanced sensitivity (80.9%) and specificity (78.9%). The 20-gene signature, derived from the features with the largest absolute regression coefficients, effectively stratified patients into high-risk and low-risk groups, with the high-risk cohort exhibiting significantly inferior overall survival (log-rank $p < 0.0001$). Incorporating both risk-associated genes (e.g., *CCNYL2*, *MDGA1*) and protective genes (e.g., *SLC16A12*, *MSH3*), the signature reflects the complex molecular landscape of ccRCC. Furthermore, biological functional enrichment analysis revealed that signature genes are predominantly enriched in cell adhesion, extracellular matrix organization, immune signaling pathways, and cell-cycle regulation—all key hallmarks of ccRCC progression and metastasis.

While the proposed 20-gene signature demonstrates strong predictive capability, several limitations should be considered when interpreting these findings. First, the model was developed and evaluated exclusively within the TCGA-KIRC dataset ($N = 281$), and the modest test set size ($n = 85$) limits subgroup evaluations; independent validation across external cohorts (e.g., GEO or ICGC datasets) remains essential to confirm cross-population generalizability. Second, because the signature was derived from FPKM RNA-sequencing data, performance across alternative quantification platforms (such as microarrays, qPCR, or NanoString) must be systematically evaluated and calibrated prior to clinical translation. Third, defining survival as a binary 5-year outcome excludes censored patient data and omits continuous time-to-event dynamics, which could be addressed in future work by extending the framework to survival-based models such as Cox proportional hazards or Random Survival Forests. Finally, the model evaluates transcriptomic expression independently of established clinicopathological parameters (such as TNM stage, Fuhrman grade, and patient age), requires *in vitro* and *in vivo* functional validation to elucidate exact disease mechanisms, and remains to be tested on non-clear cell RCC histologies.

In conclusion, we present a robust and interpretable 20-gene signature that accurately predicts 5-year survival in ccRCC patients by capturing critical biological processes underlying tumor progression. The model holds substantial potential for clinical risk stratification, personalized treatment planning, and patient selection in clinical trials. With further multi-cohort external validation, platform calibration, and prospective clinical evaluation, this transcriptomic signature could be translated into a valuable decision-support tool to improve the personalized management of patients with clear cell renal cell carcinoma.

References

1. Wéber, A., Vignat, J., Shah, R., Morgan, E., Laversanne, M., Nagy, P., ... & Znaor, A. (2024). Global burden of bladder cancer mortality in 2020 and 2040 according to GLOBOCAN estimates. World journal of urology, 42(1), 237.
2. Linehan, W. M., & Ricketts, C. J. (2019). The Cancer Genome Atlas of renal cell carcinoma: findings and clinical implications. Nature Reviews Urology, 16(9), 539-552.
3. Ljungberg, B., Bensalah, K., Canfield, S., Dabestani, S., Hofmann, F., Hora, M., ... & Bex, A. (2015). EAU guidelines on renal cell carcinoma: 2014 update. European urology, 67(5), 913-924.
4. Terrematte, P., Andrade, D. S., Justino, J., Stransky, B., de Araújo, D. S. A., & Dória Neto, A. D. (2022). A novel machine learning 13-gene signature: Improving risk analysis and survival prediction for clear cell renal cell carcinoma patients. Cancers, 14(9), 2111.

5. Takahashi, M., Rhodes, D. R., Furge, K. A., Kanayama, H. O., Kagawa, S., Haab, B. B., & Teh, B. T. (2001). Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. Proceedings of the National Academy of Sciences, 98(17), 9754-9759.
6. Cancer Genome Atlas Research Network Analysis working group: Baylor College of Medicine Creighton Chad J. 1 2 Morgan Margaret 1 Gunaratne Preethi H. 1 3 Wheeler David A. 1 Gibbs Richard A. 1, BC Cancer Agency Gordon Robertson A. 4 Chu Andy 4, Broad Institute Beroukhim Rameen 5 6 Cibulskis Kristian 6, Brigham & Women's Hospital Signoretti Sabina 7 54 59, Brown University Vandin Hsin-Ta Wu Fabio 8 8 Raphael Benjamin J. 8, University of Texas MD Anderson Cancer Center Verhaak Roel GW 9 Tamboli Pheroze 10 Torres-Garcia Wandaliz 9 Akbani Rehan 9 Weinstein John N. 9, ... & National Human Genome Research Institute Guyer Mark S. 53 Ozenberger Bradley A. 53 Sofia. Heidi J. 53. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature, 499(7456), 43-49.
7. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550.
8. Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.
9. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.
10. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery, 102(3), 148-158.
11. Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology—Head and Neck Surgery, 143(3), 331-336.
12. Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., ... & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic acids research, 50(W1), W216-W221.
13. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44-57.
14. "The Gene Ontology resource: enriching a GOld mine." Nucleic acids research 49, no. D1 (2021): D325-D334.
15. Chen, S., Jiang, L., Gao, F., Zhang, E., Wang, T., Zhang, N., ... & Zheng, J. (2022). Machine learning-based pathomics signature could act as a novel prognostic marker for patients with clear cell renal cell carcinoma: Molecular Diagnostics. British Journal of Cancer, 126(5), 771-777.
16. Li, P., Ren, H., Zhang, Y., & Zhou, Z. (2018). Fifteen-gene expression based model predicts the survival of clear cell renal cell carcinoma. Medicine, 97(33), e11839.
17. Chen, L., Xiang, Z., Chen, X., Zhu, X., & Peng, X. (2020). A seven-gene signature model predicts overall survival in kidney renal clear cell carcinoma. Hereditas, 157(1), 38.
18. Hong, Y., Lin, M., Ou, D., Huang, Z., & Shen, P. (2021). A novel ferroptosis-related 12-gene signature predicts clinical prognosis and reveals immune relevancy in clear cell renal cell carcinoma. BMC cancer, 21(1), 831.
19. Xing, Q., Zeng, T., Liu, S., Cheng, H., Ma, L., & Wang, Y. (2021). A novel 10 glycolysis-related genes signature could predict overall survival for clear cell renal cell carcinoma. BMC cancer, 21(1), 381.
20. Wang, Y., Chen, L., Ju, L., Qian, K., Wang, X., Xiao, Y., & Wang, G. (2020). Epigenetic signature predicts overall survival clear cell renal cell carcinoma. Cancer Cell International, 20(1), 564.
21. Zhan, Y., Guo, W., Zhang, Y., Wang, Q., Xu, X. J., & Zhu, L. (2015). A five-gene signature predicts prognosis in patients with kidney renal clear cell carcinoma. Computational and Mathematical Methods in Medicine, 2015(1), 842784.