Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2025, Volume 5, Issue 2, Page No: 150-166

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Evaluation of Multimodal Generative AI for Digital Pathology-Based Grading and Prognostication in Clear Cell Renal Cell Carcinoma

Mark Anderson[1]*, Lisa Wong[2], Sarah Lee[1]

[1]Department of Digital Pathology and Artificial Intelligence, College of Medicine, University of Florida, Gainesville, United States.
[2]Department of Renal Oncology and Generative AI, Faculty of Medicine, National University of Singapore, Singapore.

***E-mail** ✉ mark.anderson@ufl.edu

Received: 18 June 2025; **Revised:** 21 August 2025; **Accepted:** 28 August 2025

ABSTRACT

Clear cell renal cell carcinoma (ccRCC) represents the predominant histological variant of renal cell carcinoma (RCC) and necessitates precise pathological grading to inform accurate prognostic assessment. Nevertheless, conventional grading techniques depend primarily on the subjective evaluations of pathologists, which frequently result in considerable inconsistency. Although generative artificial intelligence (GenAI) has exhibited encouraging capabilities in medical imaging, its integration into digital pathology has received limited attention. This investigation examines the effectiveness of three leading multimodal GenAI systems—GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro—in performing ccRCC grading and forecasting patient outcomes. The analysis encompassed 499 ccRCC whole-slide images sourced from The Cancer Genome Atlas along with 349 independent external cases drawn from two separate validation cohorts. Employing a standardized prompt repetition protocol and a variance-driven stability assessment procedure, the GenAI models were directed to identify 17 distinct pathological characteristics. Stability of these extracted features was quantified via the intraclass correlation coefficient (ICC). The derived features, integrated with 3 clinical parameters, facilitated the development of grading and survival prediction models through logistic regression as well as 113 diverse machine learning techniques. Comparative benchmarking was conducted against established approaches including CellProfiler, ResNet-50, DenseNet-121, attention-based multiple instance learning (MIL), and Pathology Language and Image Pre-training, utilizing the concordance index (C-index) and area under the receiver operating characteristic curve (AUC) as primary metrics. Among the GenAI systems, Claude-3.5-Sonnet demonstrated superior performance (ICC = 0.76; micro-average AUC = 0.87), surpassing both ResNet-50 (AUC = 0.78) and attention-based MIL (AUC = 0.70). The optimal prognostic models generated by this system yielded an average C-index of 0.739 and reliably differentiated patients into high- and low-risk categories. Prominent predictive variables included tumor stage, calcification, sarcomatoid features, and vascular architecture. These findings indicate that GenAI, with particular emphasis on Claude-3.5-Sonnet, can substantially improve the precision and reproducibility of ccRCC pathological evaluation, offering significant promise for clinical deployment, particularly in environments with restricted resources.

Keywords: Clear cell renal cell carcinoma, Gen-AI, Large language models, Pathological grading, Prognostic prediction

How to Cite This Article: Anderson M, Wong L, Lee S. Evaluation of Multimodal Generative AI for Digital Pathology-Based Grading and Prognostication in Clear Cell Renal Cell Carcinoma. Interdiscip Res Med Sci Spec. 2025;5(2):150-66. https://doi.org/10.51847/o8UzmKYRdQ

Introduction

Clear cell renal cell carcinoma (ccRCC), the most frequent histological form of renal cell carcinoma (RCC), comprises roughly 70%–80% of all RCC diagnoses [1, 2]. In accordance with the International Society of Urological Pathology/World Health Organization (ISUP/WHO) guidelines, ccRCC grading follows a four-level classification scheme (grades I–IV). This grading process, which centers on microscopic evaluation of nuclear

morphology within tumor cells, serves as a well-established independent predictor of clinical outcome, especially for individuals with grade IV tumors [3, 4]. Existing grading frameworks rely heavily on individual pathologists' personal interpretations, inevitably producing variability between observers and thereby diminishing the overall reliability and standardization of results. In addition, the meticulous evaluation of extensive collections of pathological specimens is both labor-intensive and time-demanding, significantly increasing the burden on pathology professionals and raising the risk of long-term occupational burnout. These difficulties are particularly evident in underserved geographic areas lacking sufficient specialist expertise, where they hinder timely and accurate diagnosis while impairing optimal utilization of healthcare resources [5]. Hence, the availability of objective and dependable tumor grading information is critical for tailoring treatment strategies, estimating prognosis with confidence, and supporting informed clinical choices for ccRCC patients.

Over recent years, the accelerated evolution of artificial intelligence (AI) has revolutionized the processing and interpretation of complex biomedical datasets [6-12]. Within this landscape, generative artificial intelligence (GenAI) has attracted growing interest because of its advanced capacities for synthesizing and analyzing multifaceted data [12-14]. While GenAI holds substantial promise for applications in medical imaging, its practical implementation in routine pathology workflows remains largely experimental. Recent progress in AI has notably elevated standards in diagnostic imaging analysis, with marked successes reported in areas such as musculoskeletal disorder detection and chest radiograph interpretation [15, 16]. Contemporary frontier GenAI platforms, including GPT-4, Claude, and Gemini, excel in both image interpretation and linguistic reasoning, positioning them as powerful candidates for pathology support. Nevertheless, empirical research exploring their deployment in this domain remains sparse and methodologically underdeveloped. A comprehensive inquiry into GenAI applications for ccRCC grading therefore carries strong potential to advance diagnostic accuracy, uniformity, and reproducibility.

Recent scholarship concerning multimodal GenAI in computational pathology has recorded meaningful progress. Earlier iterations of GenAI were largely restricted to text-based training, limiting their capacity to interpret the rich visual complexity of histopathological slides and restricting their clinical utility. Successor multimodal models, however, have proven more adept at aiding pathologists in extracting pertinent details from reports, automating specimen classification and grading, and detecting clinically relevant morphological features [17]. Several pioneering investigations have highlighted GenAI's value in refining pathology report formats, harmonizing diagnostic language, and deriving evidence-supported recommendations from clinical documentation [18, 19]. The ongoing maturation of fusion-based multimodal GenAI architectures is creating fresh opportunities and methodological directions for both research and practical deployment in computational pathology.

State-of-the-art multimodal GenAI systems such as GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro enable direct, end-to-end extraction of morphological details and semantic insights from digitized pathology images, thereby enhancing diagnostic precision and grading consistency [20-22]. Beyond improving workflow efficiency, these models introduce transparent and interpretable AI-driven decision support tools that may fundamentally reshape research methodologies and clinical practices in digital pathology. Despite such advances, systematic appraisal of GenAI within ccRCC pathological analysis continues to expose important knowledge gaps, particularly regarding the real-world clinical value and feasibility of multimodal implementations.

This research conducts a thorough appraisal of the clinical performance of three prominent multimodal GenAI models (GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro) applied to ccRCC pathology. The central goal is to rigorously evaluate these models' capacity for accurate recognition of tumor micro-morphological characteristics, reliable pathological grading, and effective prediction of long-term survival. By leveraging automatically extracted pathological features to develop robust prognostic models and validating results across independent external cohorts, the study seeks to furnish solid evidence supporting the integration of AI into pathology decision-support systems. As the inaugural large-scale systematic examination of GenAI in ccRCC pathology, this work aims to deliver detailed scientific perspectives that can guide the clinical adoption and standardization of AI-assisted diagnostic tools. Anticipated outcomes include enhanced accuracy, operational efficiency, and inter-observer agreement in ccRCC grading, alongside strengthened capabilities for precision medicine and individualized risk stratification. Ultimately, such advancements are expected to improve patient care and therapeutic decision processes for individuals diagnosed with ccRCC.

Materials and Methods

Research data collection and quality control

This investigation involved the systematic retrieval and careful screening of 519 diagnostic whole-slide images from the The Cancer Genome Atlas (TCGA) KIRC repository. Leveraging the associated pathology reports, the team accurately extracted and recorded ccRCC grading data, retaining only those slides with unambiguous grading details. Because pathology reports are typically derived from resection specimens and may not correspond exactly with findings on diagnostic slides, cases showing notable mismatches were removed according to predefined stringent exclusion rules. After this thorough filtering, TCGA-KIRC slides containing complete disease-free survival (DFS) records—including both status and duration—were retained, yielding a final cohort of 499 high-quality diagnostic slides. These slides served as the main evaluation set for assessing the GenAI models and were also partitioned to form the training and internal validation datasets for survival modeling **(Figure 1)**. Nuclear grade assignments from the original reports were harmonized to the current ISUP/WHO grading system.

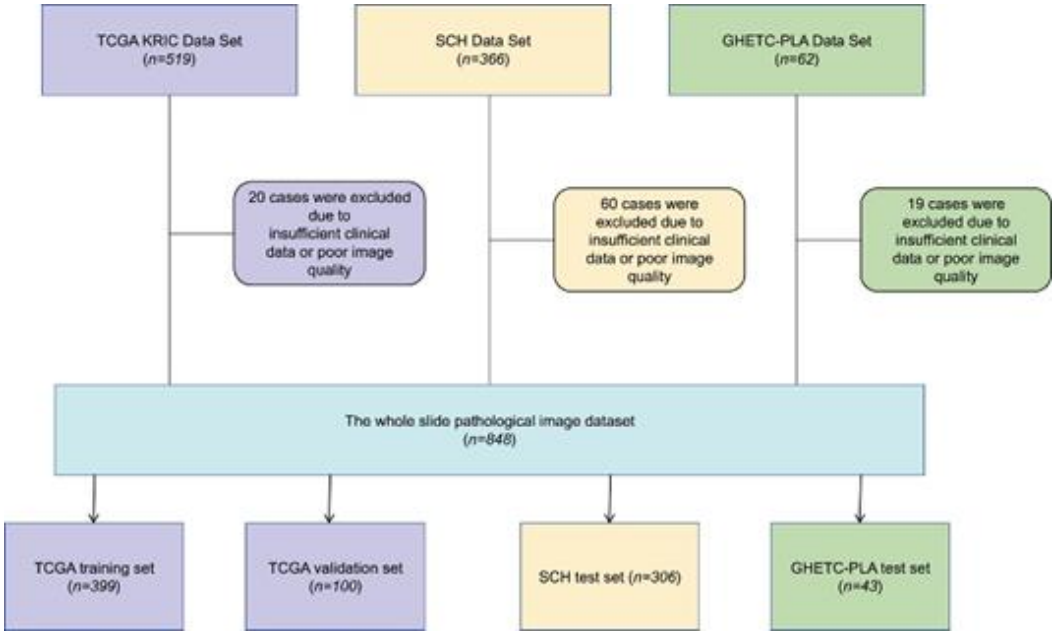


Figure 1. The patient selection procedure.

Independent external validation was performed using diagnostic slides from ccRCC patients treated surgically at Shanghai Changhai Hospital (SCH) and the General Hospital of Eastern Theater Command of the People's Liberation Army (GHETC-PLA). This external cohort comprised 349 patients in total, consisting of 306 cases (306 slides) from SCH and 43 cases (43 slides) from GHETC-PLA **(Figure 1)**. These slides functioned as completely separate test sets to objectively evaluate the survival prediction model's robustness and practical clinical performance. All procedures in this study complied with established ethical principles, and written informed consent was secured from every participant. The study protocol underwent formal review and received approval from the institutional ethics committees of both SCH and GHETC-PLA.

Demographic and clinicopathological features of the three cohorts are summarized in **Table 1**, which details the distribution of patient age, sex, tumor-node-metastasis stage, and WHO/ISUP nuclear grade.

Table 1. Demographic and clinicopathological characteristics of the cancer patients in this study.

Variable	*p*-value[b]	Cohort TCGA	Cohort SCH	Cohort GHETC-PLA	Cohort Overall
		***N* = 499**[a]	***N* = 306**[a]	***N* = 43**[a]	***N* = 848**[a]
Age, years	<0.001				
Mean (SD)		60.6 (12.2)	63.8 (11.6)	53.5 (11.1)	61.4 (12.2)
Median (Q1, Q3)		61.0 (52.0, 70.0)	65.0 (56.0, 72.0)	52.0 (48.0, 59.0)	62.0 (53.0, 71.0)
Min, max		26.0, 90.0	23.0, 92.0	30.0, 75.0	23.0, 92.0
Sex	0.005				
Female		179 (35.9%)	77 (25.2%)	16 (37.2%)	272 (32.1%)
Male		320 (64.1%)	229 (74.8%)	27 (62.8%)	576 (67.9%)
Clinical stage	<0.001				

I		251 (50.3%)	229 (74.8%)	27 (62.8%)	507 (59.8%)
II		54 (10.8%)	30 (9.8%)	6 (14.0%)	90 (10.6%)
III		117 (23.4%)	23 (7.5%)	10 (23.3%)	150 (17.7%)
IV		77 (15.4%)	24 (7.8%)	0 (0.0%)	101 (11.9%)
ISUP/WHO nuclear grade	<0.001				
1		12 (2.4%)	11 (3.6%)	3 (7.0%)	26 (3.1%)
2		215 (43.1%)	158 (51.6%)	14 (32.6%)	387 (45.6%)
3		200 (40.1%)	122 (39.9%)	21 (48.8%)	343 (40.4%)
4		72 (14.4%)	15 (4.9%)	5 (11.6%)	92 (10.8%)

- Abbreviations: GHETC-PLA, General Hospital of Eastern Theater Command of the People's Liberation Army; ISUP/WHO, International Society of Urological Pathology/World Health Organization; SCH, Shanghai Changhai Hospital; TCGA, The Cancer Genome Atlas.
- a n (%).
- b Kruskal–Wallis rank sum test; Pearson's Chi-squared test.

Prompt engineering

A detailed, structured prompt framework was first designed to enable the GenAI models to systematically identify and extract relevant features from pathological slides. Guided by the International Society of Urological Pathology/World Health Organization (ISUP/WHO) grading criteria, the researchers defined a comprehensive set of key morphological characteristics, including nuclear atypia, multinucleated giant tumor cells, rhabdoid differentiation, sarcomatoid differentiation, and several additional features of interest. The models were also explicitly directed to output ISUP/WHO nuclear grading scores consistent with official standards. In total, 17 clinically important pathological features were targeted for extraction to support a complete prognostic evaluation of kidney renal clear cell carcinoma (KIRC): nuclear grading, nest-like structure, acinar structure, tubular structure, papillary structure, coagulative necrosis, tumor necrosis, sarcomatoid differentiation, fibromyxoid stroma, calcification, ossification, vascular network, hemorrhage, degenerative scar-like component, stromal lymphocytic infiltration, rhabdoid differentiation, and ISUP/WHO nuclear grading [23-27] **(Figure 2)**.

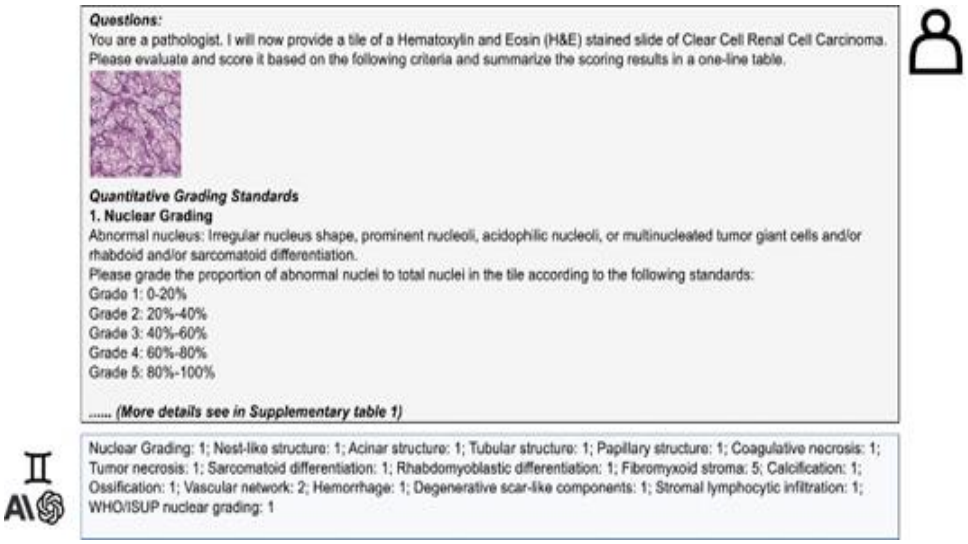


Figure 2. Presentation of the standardized process for pathological feature extraction.

Patch extraction and dataset partitioning

All whole-slide images underwent organized patch-level processing and quality assurance. An automated tissue segmentation algorithm first identified viable tissue areas on the WSIs at full resolution, generating candidate tumor-containing patches of 2048 × 2048 pixels. Three representative regions of interest (ROIs) were then selected from the tumor regions of each WSI for GenAI analysis according to specific ROI selection standards: ① well-preserved tissue architecture displaying classic ccRCC morphological traits (such as nuclear pleomorphism and nucleus-to-cytoplasm ratio) without substantial artifacts or non-tissue elements; ② even spatial coverage across the slide to represent tissue variability adequately; ③ preference for regions demonstrating grade or architectural

heterogeneity when available; and ④ selection of the largest and most characteristic patches when tumor burden was limited. These ROI choices were confirmed through independent review by two senior pathologists to guarantee biological relevance and reliability. Beyond the three primary ROIs, an additional 37 patches satisfying the same criteria were obtained from each TCGA WSI to support multivariate logistic regression modeling and ResNet-50 training.

To enable direct comparison with attention-based deep multiple instance learning (AB-MIL) [28], a broader non-ROI patch sampling approach was also implemented across all TCGA WSIs. Patches with background content exceeding 50% were discarded, and the remaining patches were retained for downstream processing. The Pathology Language and Image Pre-training (PLIP) model [29] (pre-trained on the OpenPath dataset [29]) functioned as an offline feature extractor, producing 512-dimensional feature embeddings for each patch. These embeddings were then transformed through a fully connected layer into a compatible 512-dimensional representation suitable for the MIL framework.

GenAI-based pathological feature extraction

The patches derived according to the established ROI selection protocol (excluding those designated for AB-MIL analysis) were fed into three leading multimodal GenAI platforms: GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. Guided by carefully designed prompt templates **(Figure 2)**, a systematic multi-stage procedure was followed to combine patch-level observations into comprehensive WSI-level feature representations. This involved: (1) conducting three independent evaluations of each patch with the same model, thereby generating a 3 × 17 feature matrix; (2) applying a variance-driven stability assessment with an automatic re-query protocol, where any feature displaying a standard deviation above 0.3 across the three runs prompted additional queries until acceptable stability was reached; and (3) computing the arithmetic mean across the nine available scores (three ROIs evaluated three times each) to produce a final 1 × 17 dimensional WSI-level feature vector for subsequent analysis. The complete GenAI-driven feature extraction pipeline and overall study framework are illustrated in **Figure 3**.

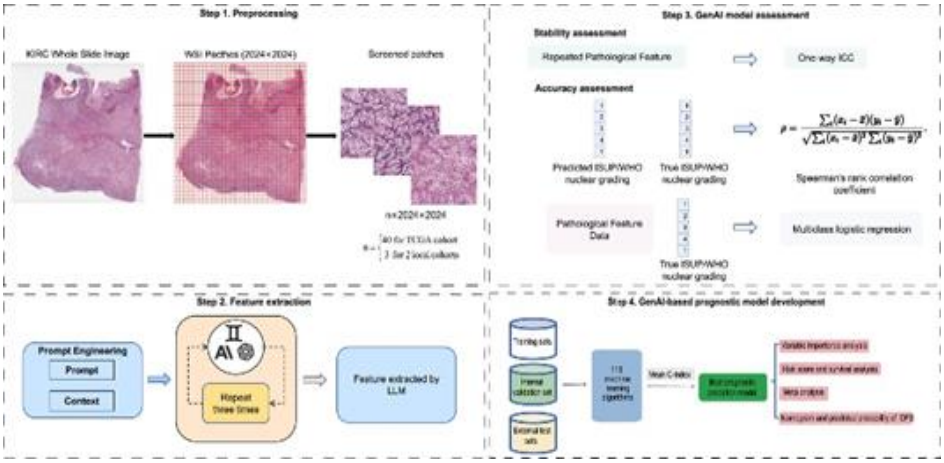


Figure 3. Flowchart Illustrating the Process of Histological Evaluation and Prognostic Model Development for ccRCC Using GenAI.

Step 1. Preprocessing: illustration of the standardized workflow for digitization of pathological slides. Step 2. Feature extraction: depiction of feature extraction from H&E-stained slides, including a systematic listing of essential pathological evaluation parameters. Step 3. GenAI model assessment: description of the thorough evaluation framework applied to three GenAI models, incorporating ICC for stability assessment, Spearman rank correlation coefficients and AUC from multinomial logistic regression for accuracy evaluation, and detailed examination of relationships between pathological features and tumor grading. Step 4. GenAI-based prognostic model development: comprehensive outline of the procedures for constructing, selecting, and analyzing prognostic models. AUC, area under the receiver operating characteristic curve; ccRCC, clear cell renal cell carcinoma; GenAI, generative artificial intelligence; ICC, intraclass correlation coefficient.

In parallel, the contribution of clinical and genomic information to grading and outcome prediction was examined through four separate experimental configurations that varied the inclusion of these data types. Comparative results revealed a consistent performance order for both nuclear grading and prognostic tasks: the combination of

GenAI pathological features with clinical and genomic variables outperformed all other groups, followed by GenAI features plus clinical variables, while GenAI features alone or combined only with genomic data performed similarly and less effectively. All observed differences were statistically meaningful. Because clinical variables are routinely accessible in daily practice whereas genomic profiling data are often incomplete or unavailable, and given the robust performance of the GenAI plus clinical features combination, this pairing was adopted for the definitive models to maximize real-world applicability. As a result, each WSI was characterized by 20 variables (17 GenAI-derived pathological features plus 3 clinical features), specifically age, gender, and tumor stage.

Experiments evaluating Macenko stain normalization indicated no meaningful changes in the 17 extracted features across the three GenAI models after normalization ($p > 0.05$). Several features remained entirely consistent, further highlighting their stability regardless of staining variations. These outcomes confirm the models' resilience to common staining differences. While stain normalization remains advisable for promoting data consistency across different laboratories, its influence on feature values and final results was minimal in this context; consequently, it was omitted from the primary workflow. The ROI selection process inherently minimized artifact interference, rendering separate artifact removal unnecessary.

To confirm that the three chosen ROIs sufficiently represented the characteristics of the full WSI, two dedicated sensitivity analyses were undertaken. Random selection-based evaluations of feature stability yielded low average standard deviations for every feature (all < 0.3), demonstrating reliable GenAI outputs when using three patches per slide across different random combinations. Assessment of how patch number influenced feature variance showed limited variation even when increasing to 10 patches per WSI, supporting the chosen strategy as an efficient balance between reliability and resource use. Additional analysis indicated that below 30 patches the performance metric (mean squared error multiplied by patch count) did not display a steady decline but fluctuated, further justifying the selection of three representative patches.

Stability of GenAI and ICC

Each GenAI model (GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro) performed three separate rounds of pathological feature extraction on every individual patch. With three ROIs per slide, the 17 feature scores for each WSI were ultimately calculated as the mean of nine measurements (3 ROIs × 3 repetitions). To evaluate the reproducibility and consistency of these scores when the identical GenAI model processed the same material repeatedly, the study applied intraclass correlation coefficient (ICC) analysis under a two-way random effects model **(Figure 3)**. This method facilitated a direct and quantitative comparison of the stability achieved by the three prominent GenAI systems in the context of ccRCC pathological feature recognition.

Accuracy of GenAI and pathological grading analysis

This study assessed the reliability of model predictions by performing linear regression and correlation analyses that compared ISUP/WHO nuclear grades generated by the GenAI systems with the corresponding grades documented in pathology reports (with TCGA reports converted to the ISUP/WHO system). These analyses enabled a detailed examination of the models' effectiveness in ccRCC grade determination **(Figure 3)**. Multivariate logistic regression was applied to predict nuclear grade for each slide, framed as a four-class classification task, utilizing the 17 GenAI-extracted pathological features along with the three clinical variables **(Figure 3)**. From the 40 ROIs obtained across all TCGA slides, 37 were designated for training while the remaining three ROIs per slide were reserved exclusively for independent validation of grading accuracy.

After conducting an extensive evaluation of stability and accuracy metrics for the three GenAI models, the research identified the strongest performer for morphological feature recognition and grading in ccRCC. Multivariate logistic regression was further used to determine the statistical relationships between the full set of 20 features and observed pathological grades, with special focus on the ability to differentiate low-grade (grade 1) from high-grade (grades 2–4) lesions. Only features showing statistically significant associations ($p < 0.05$) across grade comparisons were carried forward into final model development to maximize predictive power and improve clinical prognostic utility.

Optimal selection-based artificial intelligence model construction and validation of prognostic prediction system

The TCGA-KIRC slide patches and associated 20 features were randomly allocated to training and internal validation sets in an 8:2 proportion. Two fully independent external cohorts from SCH and GHETC-PLA served as additional test sets for robust validation. In total, 113 different machine learning algorithms were explored for

feature selection and construction of prognostic models. C-index values were calculated individually for the training set, internal validation set, and both external test sets. The model with the highest average C-index across all sets was selected as optimal **(Figure 3)**.

For the top-performing survival model, the contribution of each variable was quantified and patient-specific risk scores were computed across every dataset. Patients were subsequently divided into high-risk and low-risk categories based on these scores. Kaplan–Meier analysis evaluated survival differences between the groups, while Cox proportional hazards models estimated hazard ratios (HRs) between risk scores and disease-free survival (DFS) within each cohort. A meta-analysis was then performed to summarize the overall link between risk scores and patient outcomes across all study groups **(Figure 3)**.

Comparative evaluation of visual language models and traditional deep learning approaches in ccRCC pathological grading and prognosis prediction

The performance of GenAI models in ISUP/WHO nuclear grading of ccRCC was benchmarked against conventional deep learning methods by implementing the ResNet-50 convolutional neural network using identical training and validation partitions as those in the multivariate logistic regression analyses for direct comparability **(Figure 4a)**.

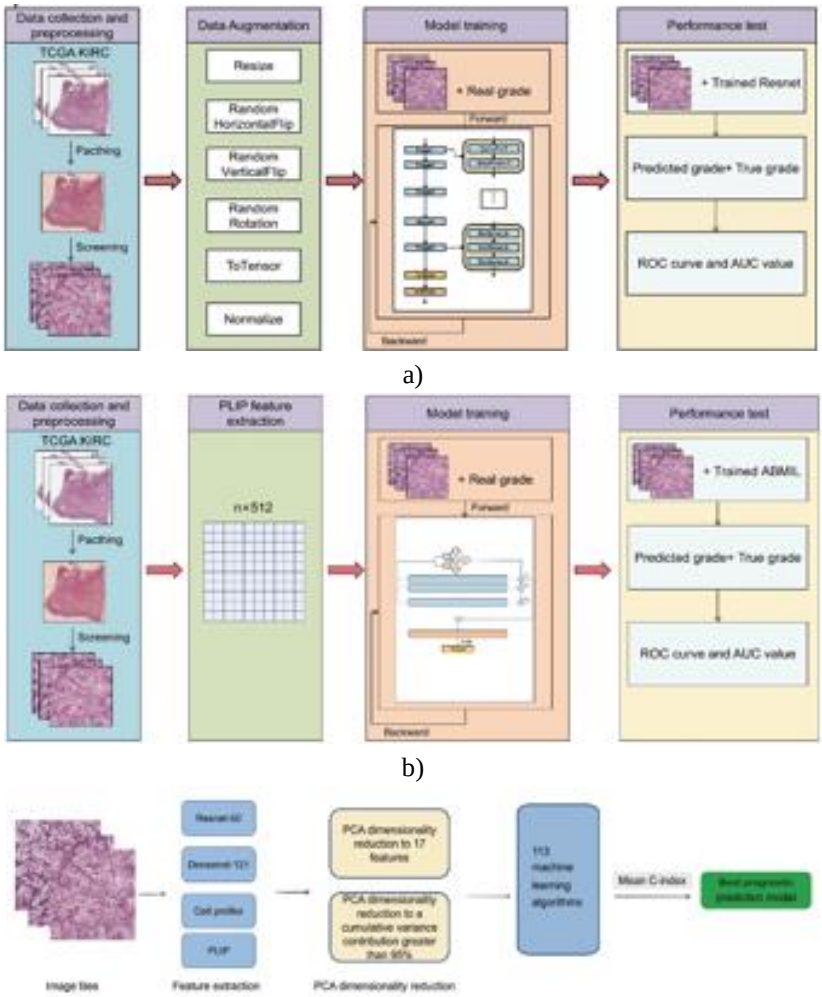


Figure 4. Comparative Assessment Between GenAI Techniques and Conventional Deep Learning Architectures as Well as Image Texture Analysis Strategies.

(a and b) Detailed depiction of the training and inference processes for pathological grade classification models based on ResNet-50 (a) and ABMIL (b), including direct performance comparison of AUC values with those obtained from GenAI multinomial logistic regression models. (c) Application of feature extraction using ResNet-

50, DenseNet-121, Cell Profiler, and the pathology foundation model PLIP, followed by two-stage PCA dimensionality reduction for enhanced feature representation. Prognostic prediction models were then developed and thoroughly benchmarked against models utilizing GenAI-derived features. ABMIL, attention-based deep multiple instance learning; AUC, area under the receiver operating characteristic curve; GenAI, generative artificial intelligence; PCA, principal component analysis; PLIP, Pathology Language and Image Pre-training.

An attention-based deep multiple instance learning (AB-MIL) model was additionally developed for nuclear grading. Dataset preparation followed the protocol described, maintaining consistency with the patient-level splits used in multivariate logistic regression **(Figure 4b)**.

To compare GenAI against traditional deep learning, image-derived feature methods, and pathology foundation models in prognostic prediction and feature representation, survival models were built from features obtained via ResNet-50, DenseNet-121, Cell Profiler, and PLIP. Their best average C-index results were contrasted with the leading model identified among the 113 machine learning approaches **(Figure 4c)**.

Statistical analysis

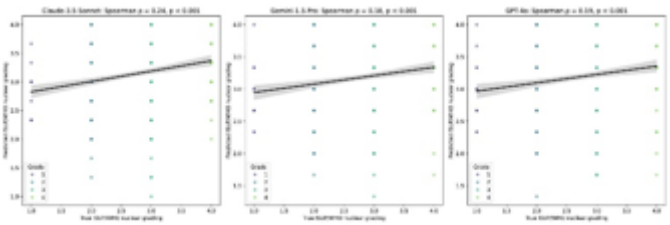
Consistency of repeated scoring by the same GenAI model on identical inputs was measured using the intraclass correlation coefficient (ICC) under a two-way random effects model. Agreement between GenAI-predicted nuclear grades and pathologist-assigned grades was quantified through area under the receiver operating characteristic curve (AUC) values and Spearman rank correlation coefficients. Statistical significance was defined as two-sided p values less than 0.05. All statistical computations were executed in R software (version 4.3.2) with relevant packages. Image preprocessing, training, classification, feature extraction, and associated procedures for ResNet-50, DenseNet-121, ABMIL, and PLIP were conducted on a 64-bit Windows environment. These tasks utilized Pytorch, OpenCV, Python (version 3.12), and the Scikit-learn library.

Results and Discussion

Stability of GenAI

The stability profiles of the three GenAI models are displayed in **Figure 5a**. While each model produced reasonably consistent scores upon repeated application, clear variations in performance were apparent. Claude-3.5-Sonnet attained the highest level of reproducibility, recording an ICC of 0.76 that signified excellent reliability. GPT-4o achieved a closely comparable ICC of 0.75, also reflecting strong consistency. Gemini-1.5-Pro registered an ICC of 0.74, which, although within a tolerable range, indicated noticeably more fluctuation than the other two. These observations highlight Claude-3.5-Sonnet as the most stable option for processing identical ccRCC samples, in contrast to Gemini-1.5-Pro, which showed comparatively lower consistency. Stability therefore emerges as a vital criterion when choosing an optimal GenAI platform for ccRCC pathological interpretation.

a)



b)

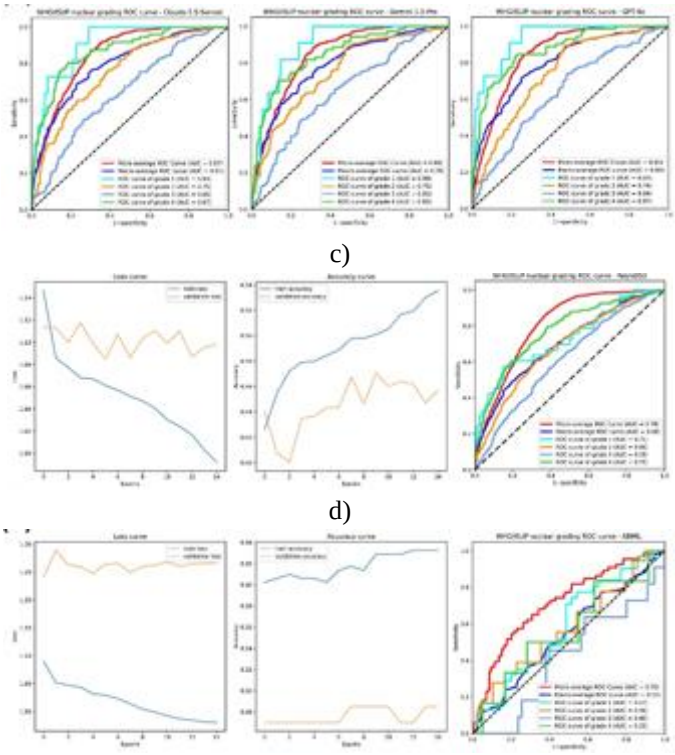


Figure 5. Evaluation of Predictive Performance for Three Leading GenAI Models (GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro) in Pathological Grade Classification of ccRCC.

(a) Heatmap representation of intraclass correlation coefficients (ICCs) for the three GenAI models across various pathological feature dimensions. Colors transition from white, denoting lower ICC values, to deep red for higher ICC values, with annotations indicating levels of statistical significance. (b) Scatter plot demonstrating the quantitative association between ISUP/WHO nuclear grades predicted by GenAI and the corresponding ground-truth grades, measured via Spearman rank correlation coefficients (R values). (c) Receiver operating characteristic (ROC) curve results for each model's ability to predict renal cell carcinoma grades, including micro-average area under the curve (AUC), macro-average AUC, and category-specific AUC values for the four grading levels. (d) In-depth summary of the ResNet-50 deep learning model training for pathological grade classification. The configuration included cross-entropy loss function, Adam optimizer at a learning rate of $3 \times 10{-}5$, batch size of 128, and 15 epochs. Loss curves highlight potential overfitting tendencies, accuracy curves track validation set performance over time, and ROC analysis reports micro-average AUC, macro-average AUC, and individual AUCs for each grade category. (e) In-depth summary of the ABMIL model training for pathological grade classification. The configuration included cross-entropy loss function, Adam optimizer at a learning rate of $2 \times 10{-}4$ with weight decay of $1 \times 10{-}5$ and cosine annealing for learning rate scheduling, batch size of 1, and 15 epochs. Loss curves highlight potential overfitting tendencies, accuracy curves track validation set performance over time, and ROC analysis reports micro-average AUC, macro-average AUC, and individual AUCs for each grade category. ABMIL, attention-based deep multiple instance learning; AUC, area under the receiver operating characteristic curve; ccRCC, clear cell renal cell carcinoma; GenAI, generative artificial intelligence; ICCs, intraclass correlation coefficients; ISUP/WHO, International Society of Urological Pathology/World Health Organization; ROC, receiver operating characteristic.

Accuracy of GenAI

Diagnostic effectiveness of the three prominent GenAI models (GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro) was examined in detail using ccRCC slides. **Figure 5b** reveals that feature scores from all models correlated significantly and positively with the adjusted pathology reports based on the ISUP/WHO grading framework ($p < 0.001$). Claude-3.5-Sonnet yielded the strongest association, with a Spearman coefficient of 0.24, compared with 0.19 for GPT-4o and 0.18 for Gemini-1.5-Pro.

Multivariate logistic regression for grade prediction showed Claude-3.5-Sonnet achieving the highest micro-average AUC of 0.87, narrowly ahead of Gemini-1.5-Pro and GPT-4o (both 0.85) **(Figure 5c)**. The model excelled particularly in recognizing grade 1 lesions (AUC = 0.94). Although its separation of grades 2 and 3 was less pronounced, the overall superior metrics positioned it as the leading choice for grading.

Gemini-1.5-Pro reached a micro-average AUC of 0.85 but displayed notable difficulties with grades 2 and 3 (AUCs 0.75 and 0.65). GPT-4o handled grades 1 and 4 effectively (AUCs 0.93 and 0.87) yet performed poorly on the intermediate grades (AUCs 0.74 and 0.64), leading to an overall micro-average AUC of 0.80.

The models generally succeeded in separating the most extreme risk groups, especially grade 1, but consistently struggled with the intermediate categories of grades 2 and 3. This common shortcoming points to an area for targeted development in future work.

Results for the traditional deep learning approaches ResNet-50 and ABMIL appear in **Figures 5d and 5e**. Both showed clear signs of overfitting throughout training. ResNet-50 attained a micro-average AUC of only 0.78 — markedly below the GenAI systems — and its per-grade AUC values lagged behind those of Claude-3.5-Sonnet. ABMIL underperformed ResNet-50 on all major metrics, including micro-average AUC, macro-average AUC, and individual grade assessments.

After thorough review of both accuracy and stability indicators, Claude-3.5-Sonnet was designated the best-performing GenAI model for histological evaluation in this ccRCC study **(Figure 6a)**. Multivariate logistic regression analysis indicated that every one of the 17 features obtained from this model, in combination with the three clinical variables, held statistically significant associations with pathological grade in at least one comparison **(Figure 6b)**. Features such as tubular structure (OR [grade 4 vs. grade 1] = 0.72, 95% CI: 0.67–0.76, $p < 0.001$), hemorrhage (OR [grade 4 vs. grade 1] = 0.78, 95% CI: 0.75–0.81, $p < 0.001$), and nested architecture (OR [grade 4 vs. grade 1] = 0.91, 95% CI: 0.89–0.92, $p < 0.001$; OR [grade 3 vs. grade 1] = 0.98, 95% CI: 0.98–0.99, $p < 0.001$) displayed significant negative links to higher grades, whereas the other relevant features showed positive associations **(Figure 6b)**.

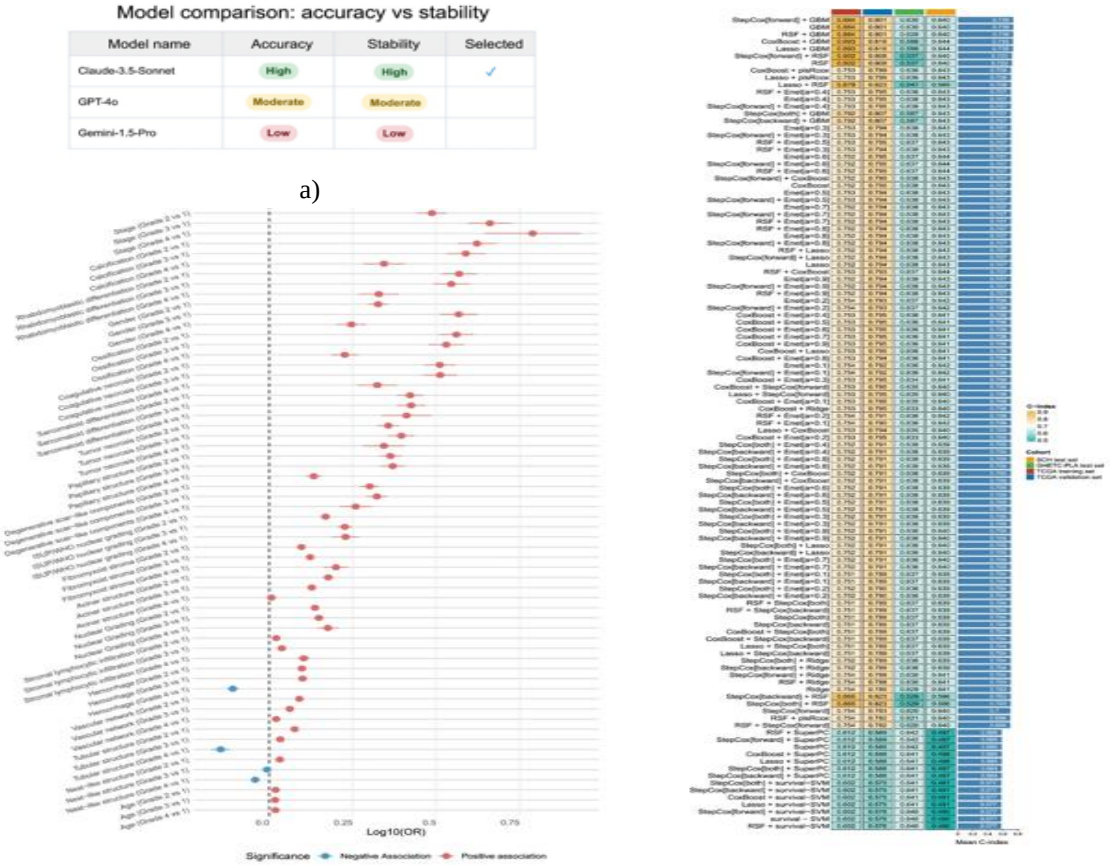


Figure 6. Establishment of Machine Learning Prognostic Models for ccRCC Using Pathological Features Derived from the Optimal GenAI System.

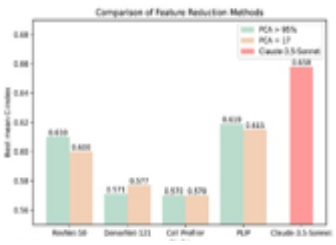
(a) Determination of the superior generative artificial intelligence model through evaluation of accuracy and stability indicators. Comprehensive comparison identified Claude-3.5-Sonnet as the most effective model. (b) Forest plot summarizing outcomes from multivariate logistic regression on histological features. The plot shows log odds ratios (ORs) for the following contrasts: ISUP/WHO grade 2 relative to grade 1, grade 3 relative to grade 1, and grade 4 relative to grade 1. (c) Heatmap evaluating the prognostic performance of 113 machine learning algorithms for disease-free survival (DFS). Blue bars denote the mean concordance index (C-index) for each model in the training set, validation set, and two external test sets. The models are ordered from highest to lowest based on their average C-index across cohorts. ccRCC, clear cell renal cell carcinoma; DFS, disease-free survival; GenAI, generative artificial intelligence; ISUP/WHO, International Society of Urological Pathology/World Health Organization; ORs, odds ratios.

Development of a machine learning prognostic model based on optimal GenAI-extracted histological feature scores

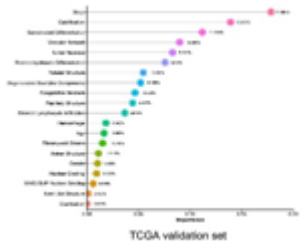
Once Claude-3.5-Sonnet was established as the best-performing GenAI system, the study divided the data into training, internal validation, and two separate external test sets following the protocol described. Features that reached statistical significance were fed into 113 different machine learning algorithms, using disease-free survival (DFS) as the main outcome. The C-index values for these models are shown in **Figure 6c**. Comprehensive evaluation identified the StepCox[forward] + Gradient Boosting Machine (GBM) combination as the top performer, with an average C-index of 0.739, making it the selected final prognostic model.

A methodical machine learning process was applied to produce trustworthy outcomes. Taking the optimal StepCox[forward] + GBM model as an example, hyperparameter optimization for GBM was carried out through grid search, testing combinations of n.trees ranging from 1000 to 4000, interaction.depth from 1 to 7, and shrinkage from 0.1 to 0.001. The best settings (n.trees = 2455, interaction.depth = 7, shrinkage = 0.001) were chosen via 10-fold cross-validation. Learning curves verified proper model convergence without overfitting, as training and validation error lines remained suitably separated, indicating solid generalization potential. The selection of the StepCox[forward] + GBM model involved initial variable screening with StepCox[forward], followed by risk prediction using GBM with Cox loss. C-index performance was then assessed across the training set and three validation groups. This model delivered the highest average C-index and was therefore adopted.

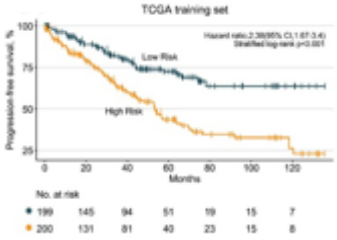
Traditional feature extraction from ImageNet30-pretrained ResNet-50 [30], DenseNet-121 [31], Cell Profiler [32], and PLIP [29] produced far more variables than the number of available samples. Two principal component analysis reduction methods were therefore applied: one retaining components until cumulative explained variance slightly surpassed 95%, and another limiting the total to 20 features to align with the GenAI plus clinical variable set. Applying the same data division strategy, 113 DFS-oriented prognostic models were developed and tested. Results in **Figure 7a** demonstrate that the best C-index values from these conventional approaches remained lower than those achieved with features from Claude-3.5-Sonnet combined with clinical information.



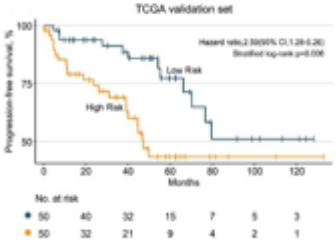
a)



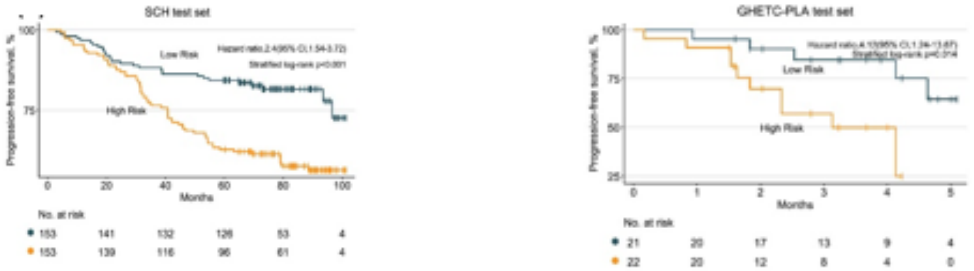
b)



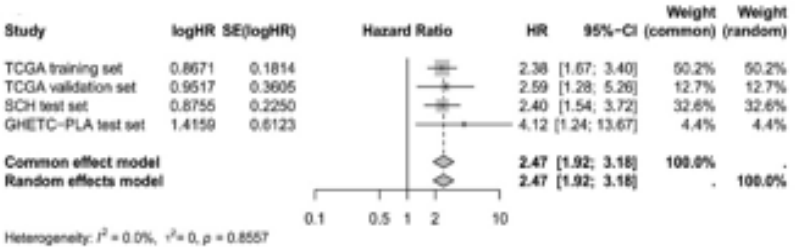
c)



d)



e) f)



g)

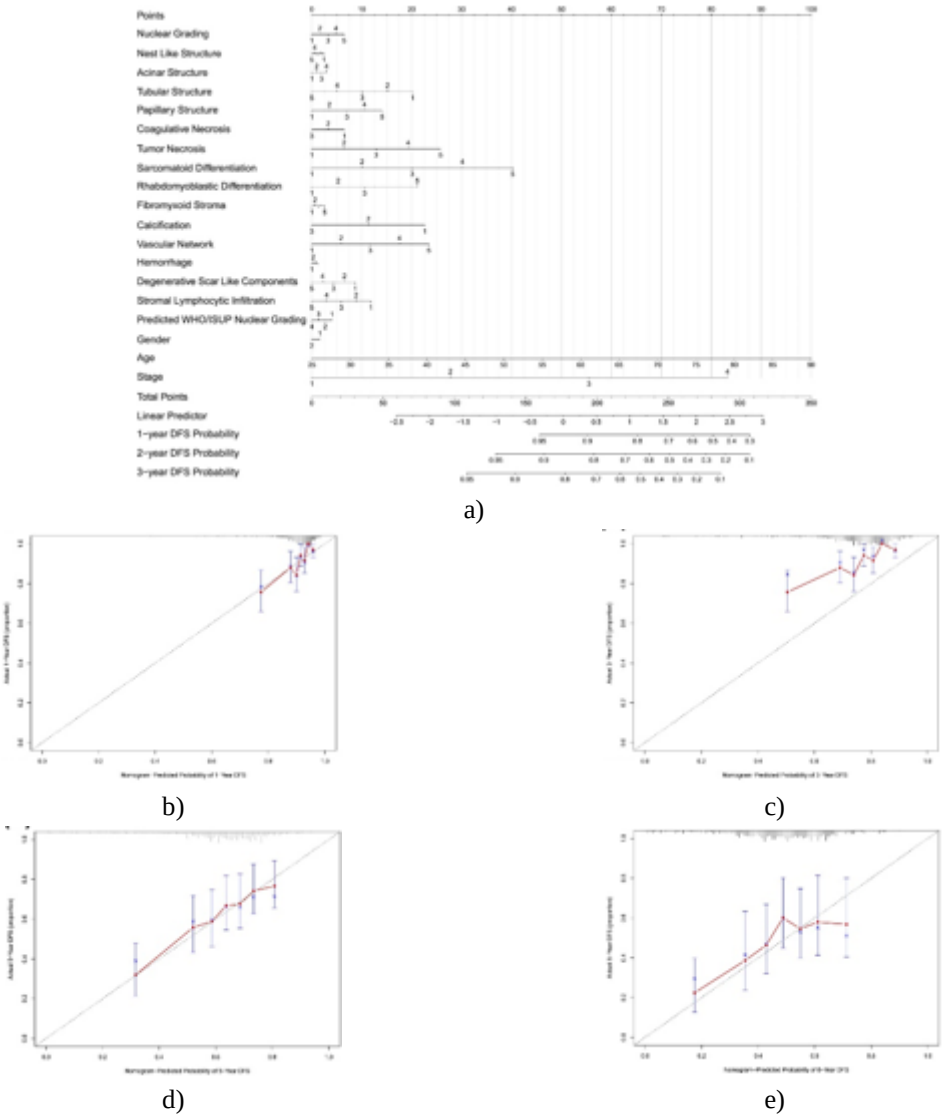
Figure 7. Overall Performance Evaluation of Refined Machine Learning-Based Prognostic Models for ccRCC Using Pathological Features.

(a) Comparative analysis of the effectiveness of prognostic model development employing features obtained via Resnet-50, DenseNet-121, Cell Profiler, and PLIP against the best-performing GenAI feature extraction strategy. (b) Assessment of the relative contribution of each predictor in the refined prognostic model (StepCox[forward] + GBM), ordered from highest to lowest impact on model predictions. (c–f) Kaplan–Meier plots illustrating disease-free survival (DFS) stratified by prognostic risk scores in the training cohort (c), internal validation cohort (d), and two separate external validation cohorts (e and f). Patients were categorized into high- and low-risk groups according to the median risk score threshold. (g) Forest plot summarizing hazard ratios (HRs) comparing high-risk and low-risk groups across the different cohorts. This plot displays the HRs along with their 95% confidence intervals (CIs) for survival outcomes between the two risk strata. Analyses were conducted on four independent datasets: the training set, validation set, and two external test sets. ccRCC, clear cell renal cell carcinoma; CI, confidence intervals; DFS, disease-free survival; GBM, Gradient Boosting Machine; GenAI, generative artificial intelligence; HRs, hazard ratios; PLIP, Pathology Language and Image Pre-training.

The contribution of each of the 20 variables in the best model is illustrated in **Figure 7b**. The most influential predictors were stage (17.836%), calcification (13.916%), sarcomatoid differentiation (11.184%), and vascular network (8.984%). Ossification contributed a weight of zero, as Claude-3.5-Sonnet detected no instances of this feature in any patches, leading to a constant score of 1 for all samples.

Risk scores generated by the final model were used to categorize patients into high-risk and low-risk groups using the median cutoff across training, validation, and external test sets. Kaplan–Meier curves confirmed significantly inferior DFS for the high-risk group in all cohorts (Log-rank $p < 0.05$) **(Figures 7c–7f)**.

Pooled meta-analysis yielded a hazard ratio of 2.47 (95% CI 1.92–3.18) for both fixed- and random-effects models, with no detectable heterogeneity ($I^2 = 0\%$, $\tau^2 = 0$, $p = 0.8557$). Cohort-specific HRs were 2.38 (training), 2.59 (validation), 2.40, and 4.12 (external tests), carrying weights of 50.2%, 12.7%, 32.6%, and 4.4%. These findings confirm substantially elevated risk in the high-risk category **(Figure 7g)**. For practical clinical application, a nomogram was created from the TCGA data to estimate individual DFS probabilities **(Figure 8a)**. Calibration plots showed strong alignment between predicted and actual outcomes, especially at 5 and 8 years, with curves closely tracking the ideal 45-degree line **(Figures 8b–8e)**.



a)

b) c)

d) e)

Figure 8. Construction of a Nomogram Using Pathological Features Derived from Gen-AI and Validation of Its Predictive Accuracy.

(a) Nomogram designed to estimate the probability of disease-free survival (DFS) in patients according to key pathological features. (b–e) Calibration plots of the nomogram showing the degree of concordance between model-predicted and observed DFS probabilities at the 1-, 3-, 5-, and 8-year time points. The performance of the nomogram is illustrated by the proximity of the curves to the ideal 45° reference line. DFS, disease-free survival; GenAI, generative artificial intelligence.

This investigation performed a comprehensive evaluation of the potential for clinical deployment of vision-enabled generative AI (GenAI) models in the identification of histopathological features and quantitative assessment of clear cell renal cell carcinoma (ccRCC). The results showed that the three tested GenAI models—GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro—all attained strong accuracy when interpreting pathological images. Claude-3.5-Sonnet achieved the highest performance with respect to both feature recognition precision and consistency of analysis. Using the 17 principal pathological features extracted by Claude-3.5-Sonnet along with three clinical variables, the study constructed reliable models for pathological grading and prognostic prediction in ccRCC. These models provided effective risk stratification in both internal and external validation sets, clearly exceeding the capabilities of conventional deep learning methods, texture-based image analysis, and pathology foundation models for feature extraction.

Of particular importance, the study identified vascular network characteristics as a key prognostic indicator in ccRCC. A standardized scoring system based on the relative area occupied by the vascular network (scored from 1 to 5) revealed that elevated vascular density (Grades 4–5, representing more than 10% of the area) correlated

strongly with adverse outcomes and disease advancement. This association may arise from several biological processes: vessels classified as Grade 4–5 exhibit leakage and irregular structure that worsen regional hypoxia, thereby driving metastatic spread through the hypoxia-inducible factor-α/vascular endothelial growth factor positive feedback loop [33, 34]; hypoxic conditions promote glycolysis, resulting in lactic acid buildup that lowers extracellular pH and triggers MMP-9-mediated basement membrane breakdown [35]; and the resulting dense vascular architecture restricts T-cell access, establishing an immunosuppressive tumor environment [36].

Standard histopathological reporting faces intrinsic constraints in terms of precise spatial mapping and quantitative evaluation of tumor characteristics, which complicates the correlation of microscopic findings with specific anatomical regions [37]. Such reports also tend to use subjective, non-standardized descriptive language that lacks objective quantitative measures. To overcome these shortcomings, the present study introduced structured prompting strategies and a detailed quantitative scoring system covering multiple critical pathological features. This quantitative framework markedly reduced the drawbacks of conventional qualitative reporting. Moreover, the implementation of multiple independent queries, variance-driven stability checks with re-querying, and feature aggregation techniques developed in this work has improved the overall robustness, dependability, and practical value of GenAI for ccRCC pathological evaluation. However, the patch-based analytical approach utilized here can result in the omission of important contextual information at the tissue level. This is especially relevant for evaluations of overall tissue architecture and tumor margins, where localized processing may generate bias. Subsequent studies should therefore examine multi-scale whole slide image (WSI) models to enhance spatial coherence and analytical transparency.

Conventional computer vision systems generally rely on CNN-based architectures or other dedicated deep learning techniques [37]. Their development demands considerable computational power and extensive labeled datasets, thereby raising both development and operational expenses [37]. GenAI models, by contrast, leverage their widespread global usage—including in low- and middle-income countries—to provide superior cost-efficiency and better resource utilization [38]. Even so, hardware demands (such as GPUs and sufficient memory) for large-model inference may still limit adoption in under-resourced environments. Strategies such as model compression, edge computing, and cloud collaboration should be pursued to enable sustainable and broadly accessible GenAI implementation [39]. Although the TCGA dataset is large and openly available, it may not reflect the full diversity of clinical populations encountered in practice. Institutional differences in staining methods and scanning devices can also cause domain shifts. Future work will therefore involve training and testing on larger, multi-center datasets to improve model robustness across varied settings.

While GenAI exhibits strong capabilities in pathological image interpretation, these systems remain vulnerable to generative biases, notably “hallucinations,” where the model produces descriptions that do not correspond to the provided image [40]. In the current analysis, an unexpected statistically significant negative association emerged between hemorrhage severity and tumor grade, likely reflecting limitations in the model’s feature detection processes. Such inaccuracies risk misleading clinicians and generating inappropriate diagnostic or therapeutic recommendations. To support safe clinical translation, a human–AI collaborative framework is recommended: the GenAI system first performs quantitative feature analysis and risk scoring, followed by pathologist review and final validation of the report [41]. Deployment will also require careful attention to ethical and legal issues. These include clarifying accountability for diagnostic errors when AI contributes to decision-making, ensuring compliance with data protection standards such as GDPR or HIPAA through anonymization and secure storage, and guaranteeing interpretability and auditability of model outputs for dispute resolution [42]. Regulatory approval from authorities such as the FDA or through CE marking will be essential for clinical auxiliary use. This necessitates rigorous validation protocols covering data diversity, multi-center testing, and pilot clinical trials, alongside quantitative risk evaluations to confirm real-world safety and performance.

In summary, this study thoroughly examined both the strengths and limitations of GenAI models for pathological grading and feature analysis in ccRCC. The evaluation was restricted to three major models; future efforts should include a more diverse array of architectures. Additional constraints involve the modest cohort size and possible biases arising from patch extraction and manual image selection procedures. Although inconsistent cases were removed, variability in slide quality across clinical samples may still have affected the results.

Conclusion

This study conducted a systematic examination of the clinical value of three prominent vision-capable GenAI models for detailed pathological evaluation of ccRCC. All three models displayed high accuracy in grading, with Claude-3.5-Sonnet performing best overall. Critical histological features identified by these models enabled the creation of integrated prognostic tools employing 113 machine learning algorithms, which maintained strong predictive performance in both internal and external cohorts. The results affirm the substantial clinical promise of AI-assisted pathology for diagnosis and outcome prediction, particularly for overcoming staffing shortages and expanding access in resource-limited healthcare environments. As the first systematic investigation of GenAI feasibility in ccRCC, this work supplies essential theoretical support and practical direction for precision kidney cancer management and the broader advancement of AI-augmented diagnostic pathology.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO classification of tumours of the urinary system and male genital organs—Part A: renal, penile, and testicular tumours. Eur Urol. 2016;70(1):93-105. doi:10.1016/j.eururo.2016.02.029
2. Jiang A, Li J, He Z, Liu Y, Qiao K, Fang Y, et al. Renal cancer: signaling pathways and advances in targeted therapies. MedComm. 2024;5(8):e676. doi:10.1002/mco2.676
3. Nishie A, Takayama Y, Asayama Y, Ishigami K, Ushijima Y, Okamoto D, et al. Amide proton transfer imaging can predict tumor grade in rectal cancer. Magn Reson Imaging. 2018;51:96-103. doi:10.1016/j.mri.2018.04.017
4. Crane A, Suk-Ouichai C, Campbell JA, Caraballo ER, Aguilar Palacios D, Tanaka H, et al. Imprudent utilization of partial nephrectomy. Urology. 2018;117:22-6. doi:10.1016/j.urology.2017.12.009
5. Malik S, Zaheer S. ChatGPT as an aid for pathological diagnosis of cancer. Pathol Res Pract. 2024;253:154989. doi:10.1016/j.prp.2023.154989
6. Lin A, Zhu L, Mou W, Yuan Z, Cheng Q, Jiang A, et al. Advancing generative artificial intelligence in medicine: Recommendations for standardized evaluation. Int J Surg. 2024;110(8):4547-51.
7. Omar M, Ullanat V, Loda M, Marchionni L, Umeton R. Chatgpt for digital pathology research. Lancet Digit Health. 2024;6(8):e595-600.
8. Zhu L, Lai Y, Mou W, Zhang H, Lin A, Qi C, et al. Chatgpt's ability to generate realistic experimental images poses a new challenge to academic integrity. J Hematol Oncol. 2024;17:27.
9. Zhu L, Mou W, Hong C, Yang T, Lai Y, Qi C, et al. The evaluation of generative AI should include repetition to assess stability. JMIR Mhealth Uhealth. 2024;12:e57978.
10. Lin A, Qi C, Li M, Guan R, Imyanitov EN, Mitiushkina NV, et al. Deep learning analysis of the adipose tissue and the prediction of prognosis in colorectal cancer. Front Nutr. 2022;9:869263.
11. Zhu L, Mou W, Wu K, Lai Y, Lin A, Yang T, et al. Multimodal ChatGPT-4V for electrocardiogram interpretation: Promise and limitations. J Med Internet Res. 2024;26:e54607.
12. Zhu L, Mou W, Lai Y, Chen J, Lin S, Xu L, et al. Step into the era of large multimodal models: A pilot study on ChatGPT-4V(ision)'s ability to interpret radiological images. Int J Surg. 2024;110(7):4096-102.
13. Chen J, Zhu L, Mou W, Zeng D, Qi C, Liu Z, et al. Stager checklist: Standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability. iMetaOmics. 2024;1:e7.
14. Akhtar ZB. Unveiling the evolution of generative AI (GAI): A comprehensive and investigative analysis toward LLM models (2021–2024) and beyond. J Electr Syst Inf Technol. 2024;11:22.
15. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. Chatgpt's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol. 2025;35(1):506-16.

16. Zhou Y, Ong H, Kennedy P, Wu CC, Kazam J, Hentel K, et al. Evaluating GPT-4V (GPT-4 with vision) on detection of radiologic findings on chest radiographs. Radiology. 2024;311:e233270.
17. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med. 2024;7(1):106.
18. Apornvirat S, Thinpanja W, Damrongkiet K, Benjakul N, Laohawetwanit T. Comparing customized ChatGPT and pathology residents in histopathologic description and diagnosis of common diseases. Ann Diagn Pathol. 2024;73:152359.
19. Steimetz E, Minkowitz J, Gabutan EC, Ngichabe J, Attia H, Hershkop M, et al. Use of artificial intelligence chatbots in interpretation of pathology reports. JAMA Netw Open. 2024;7(5):e2412767.
20. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Sonoda Y, Nakamura Y, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. Jpn J Radiol. 2024;42(12):1399-402.
21. Gemini Team Google. Gemini: A family of highly capable multimodal models. 2024.
22. Shahriar S, Lund BD, Mannuru NR, Arshad MA, Hayawi K, Bevara RVK, et al. Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. Appl Sci. 2024;14:7782.
23. Andreiana BC, Stepan AE, Mărgăritescu C, Al Khatib AM, Florescu MM, Ciurea RN, et al. Histopathological prognostic factors in clear cell renal cell carcinoma. Curr Health Sci J. 2018;44(3):201-5.
24. Schnuelle P. Renal biopsy for diagnosis in kidney disease: Indication, technique, and safety. J Clin Med. 2023;12:6424.
25. Motzer RJ, Penkov K, Haanen J, Rini B, Albiges L, Campbell MT, et al. Avelumab plus axitinib versus sunitinib for advanced renal-cell carcinoma. N Engl J Med. 2019;380(12):1103-15.
26. Zarrabi K, Walzer E, Zibelman M. Immune checkpoint inhibition in advanced non-clear cell renal cell carcinoma: Leveraging success from clear cell histology into new opportunities. Cancers (Basel). 2021;13(15):3652.
27. Kristiansen G, Delahunt B, Srigley JR, Lüders C, Lunkenheimer JM, Gevensleben H, et al. Vancouver classification of renal tumors: Recommendations of the 2012 consensus conference of the International Society of Urological Pathology (ISUP). Pathologe. 2015;36(3):310-6.
28. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. Proc 35th Int Conf Mach Learn. 2018:2127-36.
29. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29(9):2307-16.
30. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conf Comput Vis Pattern Recogn. IEEE; 2009:248-55.
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition.in 2016 IEEE Conf Comput Vis Pattern Recogn. IEEE; Las Vegas, NV, USA 2016. p. 770–8.
32. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conf Comput Vis Pattern Recogn. IEEE; 2017. p. 2261-9.
33. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. Nat Rev Dis Primers. 2017;3:17009.
34. Yin CL, Ma YJ. The regulatory mechanism of hypoxia-inducible factor 1 and its clinical significance. Curr Mol Pharmacol. 2024;17:e18761429266116.
35. Goveia J, Rohlenova K, Taverna F, Treps L, Conradi LC, Pircher A, et al. An integrated gene expression landscape profiling approach to identify lung tumor endothelial cell heterogeneity and angiogenic candidates. Cancer Cell. 2020;37(1):21-36.e13.
36. Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, Bacac M, et al. An immune atlas of clear cell renal cell carcinoma. Cell. 2017;169(4):736-49.e18.
37. Bahadir CD, Omar M, Rosenthal J, Marchionni L, Liechty B, Pisapia DJ, et al. Artificial intelligence applications in histopathology. Nat Rev Electr Eng. 2024;1:93-108.
38. Zhang Y, Liu H, Sheng B, Tham YC, Ji H. Preliminary fatty liver disease grading using general-purpose online large language models: ChatGPT-4 or Bard? J Hepatol. 2024;80(6):e279-81.
39. Liu J, Du Y, Yang K, Wang Y, Hu X, Wang Z, et al. Edge-cloud collaborative computing on distributed intelligence and model optimization: a survey. 2025.

40. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1-38.
41. Steiner DF, Chen PHC, Mermel CH. Closing the translation gap: AI applications in digital pathology. Biochim Biophys Acta Rev Cancer. 2021;1875(1):188452.
42. Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. Proc Conf Fairness Accountability Transpar. 2019:279-88.