



Published in final edited form as:

Med Phys. 2025 January ; 52(1): 300–320. doi:10.1002/mp.17476.

Aggressiveness classification of clear cell renal cell carcinoma using registration-independent radiology-pathology correlation learning

Indrani Bhattacharya^{#1}, Karin Stacke^{#2,3}, Emily Chan⁴, Jeong Hoon Lee¹, Justin R. Tse¹, Tie Liang¹, James D. Brooks⁵, Geoffrey A. Sonn^{#1,5}, Mirabela Rusu^{#1,5}

¹Department of Radiology, Stanford University, Stanford, California, USA

²Sectra, Linköping, Sweden

³Department of Science and Technology, Linköping University, Linköping, Sweden

⁴Department of Pathology, Stanford University, Stanford, California, USA

⁵Department of Urology, Stanford University, Stanford, California, USA

These authors contributed equally to this work.

Abstract

Background: Renal cell carcinoma (RCC) is a common cancer that varies in clinical behavior. Clear cell RCC (ccRCC) is the most common RCC subtype, with both aggressive and indolent manifestations. Indolent ccRCC is often low-grade without necrosis and can be CHmonitored without treatment. Aggressive ccRCC is often high-grade and can cause metastasis and death if not promptly detected and treated. While most RCCs are detected on computed tomography (CT) scans, aggressiveness classification is based on pathology images acquired from invasive biopsy or surgery.

Purpose: CT imaging-based aggressiveness classification would be an important clinical advance, as it would facilitate non-invasive risk stratification and treatment planning. Here, we present a novel machine learning method, Correlated Feature Aggregation By Region (CorrFABR), for CT-based aggressiveness classification of ccRCC.

Methods: CorrFABR is a multimodal fusion algorithm that learns from radiology and pathology images, and clinical variables in a clinically-relevant manner. CorrFABR leverages registration-independent radiology (CT) and pathology image correlations using features from vision transformer-based foundation models to facilitate aggressiveness assessment on CT images. CorrFABR consists of three main steps: (a) *Feature aggregation* where region-level features are extracted from radiology and pathology images at widely varying image resolutions, (b) *Fusion*

Correspondence Indrani Bhattacharya and Mirabela Rusu, Department of Radiology, Stanford University, Stanford, CA 94305, USA. indrani.electrical@gmail.com and mirabela.rusu@stanford.edu.

CONFLICT OF INTEREST STATEMENT

Justin R. Tse has active or recent funding from GE HealthCare and Bayer, and consult for Intuitive Surgical Inc, ABSolutions Inc, and Ascelia Pharma AB. The other authors have no conflicts to disclose.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

where radiology features correlated with pathology features (pathology-informed CT biomarkers) are learned, and (c) *Classification* where the learned pathology-informed CT biomarkers, together with clinical variables of tumor diameter, gender, and age, are used to distinguish aggressive from indolent ccRCC using multi-layer perceptron-based classifiers. Pathology images are only required in the first two steps of CorrFABR, and are not required in the prediction module. Therefore, CorrFABR integrates information from CT images, pathology images, and clinical variables during training, but for inference, it relies solely on CT images and clinical variables, ensuring its clinical applicability. CorrFABR was trained with heterogeneous, publicly-available data from 298 ccRCC tumors (136 indolent tumors, 162 aggressive tumors) in a five-fold cross-validation setup and evaluated on an independent test set of 74 tumors with a balanced distribution of aggressive and indolent tumors. Ablation studies were performed to test the utility of each component of CorrFABR.

Results: CorrFABR outperformed the other classification methods, achieving an ROC-AUC (area under the curve) of 0.855 ± 0.0005 (95% confidence interval: 0.775, 0.947), F1-score of 0.793 ± 0.029 , sensitivity of 0.741 ± 0.058 , and specificity of 0.876 ± 0.032 in classifying ccRCC as aggressive or indolent subtypes. It was found that pathology-informed CT biomarkers learned through registration-independent correlation learning improves classification performance over using CT features alone, irrespective of the kind of features or the classification model used. Tumor diameter, gender, and age provide complementary clinical information, and integrating pathology-informed CT biomarkers with these clinical variables further improves performance.

Conclusion: CorrFABR provides a novel method for CT-based aggressiveness classification of ccRCC by enabling the identification of pathology-informed CT biomarkers, and integrating them with clinical variables. CorrFABR enables learning of these pathology-informed CT biomarkers through a novel registration-independent correlation learning module that considers unaligned radiology and pathology images at widely varying image resolutions.

Keywords

clear cell renal cell carcinoma; correlated feature learning; deep learning; machine learning; multimodal fusion; radiology-pathology fusion

1 | INTRODUCTION

In 2022, there were nearly 450 000 new kidney cancer cases diagnosed, and over 150 000 reported deaths world-wide.¹ Out of these, 90%–95% are estimated to be renal cell carcinoma (RCC).² Increased utilization of abdominal imaging has led to greatly increased detection of incidental renal masses.^{3,4} Radiological imaging, such as computed tomography (CT), plays a vital role in risk assessment and treatment planning for these renal masses, by distinguishing benign masses from RCC, and by characterizing the aggressiveness of RCC.⁵

Clear cell RCC (ccRCC) is the most common subtype ($\approx 75\%$), accountable for the majority of RCC deaths.⁶ ccRCC is asymptomatic in early stages, and 25%–30% of patients have ccRCCs that have metastasized at the time of diagnosis.^{7,8} The prognosis of ccRCC depends on the aggressiveness of the mass, with higher tumor grade and the presence of necrosis being correlated with worse clinical outcomes.^{9–11}

Clinical assessment of the aggressiveness of ccRCC based on tumor grade and necrosis is conducted using pathology images of tissue collected via biopsy or surgery. These invasive procedures are often associated with adverse side effects of bleeding, pain, and infection. Making this aggressiveness assessment on non-invasive radiology images can help decide the optimum management strategy, and prevent overtreatment and adverse side effects in patients with indolent disease.^{7,12,13} However, radiology imaging-based aggressiveness classification of ccRCC is challenging due to the subtle and overlapping radiology features of the two classes (indolent and aggressive).^{14,15} Tumor size is one such example, where larger tumors can indicate a more aggressive subtype, but it is not a conclusive feature for classification.^{12,16}

Prior machine-learning based image-analysis methods for kidney tumors focused on differentiating between malignant or benign masses,^{17–19} classifying histological subtypes,^{20–23} or distinguishing between low- and high-grade ccRCC.^{24–34} In this study, we extend the definition of aggressiveness to include both tumor grade and presence of necrosis, as prior studies show that both of these measures have prognostic value in predicting disease-specific progression-free survival.^{15,35} No prior study attempted to classify ccRCC using CT images as aggressive or indolent using machine learning, based on both tumor grade and necrosis observed on pathology images. Moreover, most prior machine learning studies used radiology images in a vacuum while classifying kidney masses. Only one study considered complementary pathology images and genetic data in addition to radiology images for prognosis prediction in ccRCC.³⁶ The study showed that the multi-domain approach has the best performance, suggesting the benefit of complementary multimodal data fusion. Yet, the requirement for all three modalities during inference limits the clinical relevance of this approach, since neither histological nor genetic data are available in the pre-operative setting.

Methods incorporating pathology image information during model training, but not requiring it during inference, have been described for prostate cancer,^{37,38} but to our knowledge have not been developed for kidney cancer. The radiology-pathology fusion approach developed for the prostate, Correlated Signature Network for Aggressive and Indolent Cancer Detection (CorrSigNIA), can detect and differentiate between indolent and aggressive prostate cancer on magnetic resonance imaging (MRI).³⁸ CorrSigNIA emphasizes disease pathology characteristics on MR images by learning MRI features that are correlated with pathology features on a perpixel level. The method relies on registered (spatially aligned) MRI and pathology images for creating these pathology-correlated radiology image biomarkers. The study showed that including pathology-correlated MRI biomarkers improved the performance of MRI-based prostate cancer detection.

However, several factors limit the applicability of CorrSigNIA to most multimodal clinical datasets, including our task of aggressiveness classification of ccRCC. The foremost of these limiting factors is CorrSigNIA's dependence on spatially aligned radiology and pathology images. Most clinical radiology-pathology datasets are unaligned, and radiology-pathology image registration is challenging and often impossible in most clinical scenarios. For instance, most pathology data is collected via biopsy, where the spatial location of the biopsy needle is not tracked or stored. Moreover, pathology samples collected from surgery

are either too soft or lack spatial information about their location in the body, making it impossible to reliably and accurately match them to the preoperative radiology images. The second limiting factor of CorrSigNIA is that it severely downsamples the pathology images for learning pixel-level correlations between registered radiology and pathology images. Radiology and pathology images vary widely in image resolutions, and such downsampling of pathology images may lead to the loss of valuable information. Third, CorrSigNIA also leverages rich pixel-level annotations of aggressive and indolent cancer on both radiology and pathology images. Such pixel-level annotations on both radiology and pathology images require specialized clinical expertise, and extensive labor and time, and is difficult to obtain for most clinical scenarios.

Our dataset of classifying ccRCC into aggressive and indolent subtypes consists of spatially unaligned radiology and diagnostic pathology images at widely varying image resolutions, and without pixel-level cancer annotations. As such, CorrSigNIA cannot be applied to this dataset as an approach for radiology-pathology image fusion and necessitates the development of registration-independent multimodal image-fusion methods.

To integrate radiology and pathology image information in a clinically relevant way, we developed a registration-independent multimodal image fusion approach, CorrFABR. CorrFABR leverages recent advances in foundation model development with common representation learning to identify pathology-informed radiology biomarkers for kidney cancer. Radiology CT features are extracted from the pre-trained vision transformer-based foundation model, DinoV2³⁹ that has demonstrated excellent performance in several computer vision tasks. The recent vision-language pathology foundation model CONCH (CONtrastive learning from Captions for Histopathology)⁴⁰ is used to select representative cancer patches from pathology whole slide images (WSI) using zero-shot classification, in the absence of expert pathologist annotations. CorrFABR also uses CONCH as a pathology feature encoder. After feature extraction, CorrFABR uses registration-independent common representation learning to learn pathology-informed CT biomarkers (i.e., CT features that are maximally correlated with pathology features). These pathology-informed CT biomarkers are then used in combination with clinical variables of tumor diameter, gender, and age to classify ccRCC tumors as aggressive or indolent. The motivation of incorporating tumor diameter as an additional feature is supported by prior studies,^{12,16} which show a correlation between tumor diameter and aggressiveness. Age and sex are risk factors for developing RCC, with men experiencing higher morbidity rates than women.⁴¹ CorrFABR does not require spatial alignment between radiology and pathology images, does not require cancer annotations on the pathology WSI, and does not require pathology images during inference, allowing its use to guide biopsy procedures or facilitate treatment planning.

Using multi-institution public data without spatial correspondences, CorrFABR was trained and evaluated on cohorts from the 2021 Kidney Tumor Segmentation Challenge⁴² and The Cancer Imaging Archive (TCIA).⁴³ The main contributions of this paper, which are encompassed by the proposed CorrFABR method, can be summarized as:

1. Development of a novel 'Feature Aggregation by Region' module that (a) leverages foundation models to extract radiology and pathology features,

(b) identifies pathology-informed radiology biomarkers through registration-independent correlated feature learning using these foundation model-derived features, and (c) enables integration of spatially unaligned radiology and pathology images with widely different image resolutions.

2. Development of a novel non-invasive method to classify ccRCC into indolent or aggressive subtypes using a clinically relevant multimodal (radiology, pathology, clinical variables) fusion approach.

We posit that our proposed CorrFABR method can be widely applicable for selecting radiology cancer biomarkers through registration-independent radiology-pathology correlation learning, and can be useful in other cancer classification/segmentation tasks as well beyond this specific ccRCC classification task.

2 | MATERIALS AND METHODS

2.1 | Dataset

We used publicly available data from the 2021 Kidney and Kidney Tumor Segmentation Challenge (KiTS21) and TCIA: The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC). All patients included in this study had confirmed ccRCC, underwent contrast-enhanced CT scans, and, in the case of TCGA-KIRC, had corresponding diagnostic pathology images from partial or radical nephrectomy. This resulted in 203 patients with CT from the KiTS21 dataset,⁴² and 169 patients with CT and pathology from the TCGA-KIRC cohort⁴⁴ (Table 1). The mean tumor diameter was 5.01 ± 3.11 cm for KiTS21 and 5.83 ± 2.89 cm for TCGA-KIRC (Table 2). Figure 1 shows example data from a sample patient.

Labels and annotations: The KiTS21 data included manually segmented regions of the kidneys and tumors. For the TCGA-KIRC data, no manual segmentations were available. Automatic segmentations of the kidneys and tumors were extracted on CT using the previously validated and pre-trained Medical Image Segmentation with Convolutional Neural Networks model.⁴⁵ These segmentations were visually inspected for quality assurance. On pathology data, tissue was segmented from background using Otsu thresholding.⁴⁶ Since the pathology WSI did not have any pathologist annotations of cancer regions, an expert pathologist at our institution outlined the extent of cancer on 20 randomly selected cases (10 aggressive, 10 indolent) to enable assessment of the foundation model CONCH⁴⁰ for cancer patch selection. These annotations were only used for assessment of the quality of patch selection using CONCH, and were not used during data selection or model training.

Each tumor was labeled as either indolent or aggressive using the pathology reports. Aggressive tumors were defined as having necrosis or a high tumor grade of 3 or 4 (Fuhrman grade⁴⁷ or International Society of Urological pathology [ISUP] grade¹¹). Indolent tumors were defined as those with low tumor grade (1–2) without necrosis.¹⁶ Tumor grade and necrosis were both included in the definition of aggressiveness since both are considered important prognostic markers for ccRCC.^{35,47–49} Out of the total 372 tumors, 199 were labeled as aggressive. Of these, 19 low-grade (grade 2) cases had the presence

of necrosis (Table 3) and were thus considered aggressive despite being low tumor grade. While ccRCC tumors may be heterogenous, with both low- and high-grade areas, the lack of pixel-wise annotations restricted labels with finer granularity.

Pre-processing: All CT volumes were cropped using a three-dimensional bounding box extracted from the lesion segmentations and resized to an in-plane dimension of 224×224 pixels. The motivation of resizing all tumors to the same in-plane dimension was to enable extraction of similar CT imaging features for all cases, irrespective of the tumor size. This makes the tumor diameter a complementary clinical feature, and the predictive value of both the imaging features and the tumor diameter could be assessed independently. Due to the variability of stain color in the public data, the pathology images were stain normalized.⁵⁰

2.2 | Developing CorrFABR

The overarching goal of this study is to classify ccRCC tumors into aggressive or indolent based on tumor grade and necrosis. In order to enable this non-invasive classification from CT scans, this study presents a method for incorporating disease characteristics from pathology images into the radiology domain without requiring spatial alignment between the two domains. CorrFABR also enables integration of clinical variables of tumor diameter, gender, and age to the classification module. An overview of the proposed method, CorrFABR, is shown in Figure 2, which consists of three steps:

1. Feature extraction and aggregation from radiology CT and pathology images
2. Fusion to learn pathology-informed CT biomarkers, that is, a combination of CT features that are maximally correlated with pathology features (denoted as CorrFeat)
3. Classification of ccRCC tumors as aggressive or indolent using these learned pathology-informed CT biomarkers and clinical variables (tumor diameter, gender, age).

During training, unaligned radiology, and pathology images originating from the same patient are used in Step 1 (Feature aggregation) and Step 2 (Fusion). For training of the model in the Step 3 (Classification), as well as during inference, pathology images are not needed. Therefore, the model only relies on pathology image data during the training phase (steps a and b), and never during inference. Each step of the method is described below.

2.2.1 | Input data—During training, for each patient, image data from the radiology and pathology domains are used. Each slice in the CT volume containing the kidney mass and the corresponding pathology WSI are used as inputs to step a, feature extraction and aggregation.

2.2.2 | CorrFABR steps

Step (a): Feature extraction and aggregation: To enable learning correlated features from unregistered radiology and pathology images with widely varying image resolutions, we propose a region-level feature extraction and aggregation module. In contrast to pixel-level correlation learning used in CorrSigNIA,³⁸ region-level correlation learning extracts and

aggregates features from “regions” denoting corresponding areas in the two modalities with similar tissue composition (e.g., aggressive ccRCC). The region in one modality does not need to be spatially aligned to the region in the other modality, but both regions need to represent tissue of similar composition. For example, the kidney tumor on the CT image, and the cancer patches from the WSI of pathology acquired from the partial or complete surgical removal of the tumor, and used diagnostically to classify this tumor as aggressive or indolent ccRCC, can be considered as corresponding regions of similar tissue composition in the two modalities. Since the kidney pathology dataset does not contain pixel-level annotations of normal tissue and cancerous tissue or pixel-level annotations of aggressive and indolent ccRCC, we used a vision-language pathology foundation model CONCH⁴⁰ in a zero-shot classification framework to select cancer patches from the WSI. CONCH was trained with a diverse dataset of 1.17 million histopathology images and corresponding images captions using contrastive learning from captions in a task-agnostic way. CONCH generates pathology image features that achieve state-of-the-art performance in diverse downstream tasks on 14 diverse datasets. Since CONCH was not trained with publicly available histopathology images, the risk of data contamination on the TCGA-KIRC dataset used in this study is minimal. In particular, in the published study, CONCH was evaluated on the TCGA RCC dataset for zero-shot classification of RCC subtypes, and demonstrated excellent performance in that classification, supporting its utility in this study.

The region-based approach is in contrast to the CorrSigNIA study where very rich pixel-level annotations for the entire prostate were available in both modalities in addition to spatial alignment, enabling pixel-level correlation learning. Our proposed region-level correlation learning would be appropriate for most medical datasets with unaligned multimodal images at varying image resolutions, and without pixel-level annotations on both modalities.

Selecting matched “regions” in radiology and pathology images: Only one whole-slide pathology image was available per patient. The pathology images contain the diagnostic slice obtained from the surgical removal of either a portion or the complete kidney mass. Pathology patches of 224×224 pixels were extracted from the tissue regions of the WSI, after stain normalization and background segmentation using Otsu thresholding. Each patch was then subjected to a zero-shot classification with CONCH using text prompts asking for the patch to be classified as either “cancer” and “normal tissue”. A CONCH-predicted normalized cancer similarity score of 0.9 was used to select representative cancer patches from the WSI. This threshold of the predicted normalized cancer similarity score was decided based on ablation studies comparing CONCH performance with an expert pathologist (see Section 3.1.1). The number of patches per WSI depends on the width and height of the WSI, and the number of cancer patches selected by CONCH.

For CT, the regions were selected as the outlined tumor area. However, only slices with at least 5% of its area containing the tumor were considered. This resulted in excluding slices at the very beginning and the end of the 3D tumor volume with very small tumor area compared with the external regions outside the tumor. If $N p_j^{CONCH}$ are the number of CONCH-selected cancer patches for case j , then the kidney tumor on CT and the $N p_j^{CONCH}$

patches from the diagnostic pathology slice were considered as matched “regions” with similar tissue composition.

Feature extraction and aggregation for radiology images: The pre-trained vision transformer-based foundation model DinoV2³⁹ was used to extract CT imaging features on a slice-level. DinoV2 was trained on 142 million natural images in a self-supervised learning setup, without any labels or annotations. Prior studies show that DinoV2 generates robust visual features that can be used for a large variety of computer vision tasks.³⁹

Feature extraction and aggregation for pathology images: For pathology images, features were extracted from the *high-res* images (pixel sizes of 0.008 – 0.032 mm/pixel) in a patch-wise manner, with a patch size of 224×224 pixels, and patches selected via CONCH⁴⁰ as described above. CONCH was also used as the pathology feature encoder.

Generating paired radiology and pathology input vectors for training the fusion module: Figure 3 illustrates how the input vectors with matched radiology and pathology representations were generated to allow training of the step (b), fusion, of the CorrFABR model. Slice-level feature representations were first generated for the resized tumor volumes of each case (Step A in Figure 3). For case j with N_{c_j} slices, each slice $i \in N_{c_j}$ has a radiology feature representation R_i of dimension $(1, FD_{rad})$. To generate a corresponding pathology representation for this slice i , 75% of CONCH-selected cancer patches for the same case ($N_{p_j}^{CONCH}$) were randomly selected, generating $N_{p_{ji}}$ patches corresponding to CT slice i for case j . Pathology feature representations were extracted on a patch-level and averaged across all the $N_{p_{ji}}$ patches (Step B in Figure 3). Randomly selecting a subset of the patches introduces some variability in the pathology vector (which otherwise would be identical for all N_{c_j} slices) to enable more robust common representation learning training. The percentage of 75 was selected based on the Positive Predictive Value of the CONCH-selected patches being cancer (see Section 3.1.1).

Step B was repeated for all the N_{c_j} slices of case j , resulting in a case-level radiology feature representation S_j^{rad} with dimension (N_{c_j}, FD_{rad}) , and a case-level pathology feature representation S_j^{path} with dimension (N_{c_j}, FD_{path}) . For CorrFABR that uses DinoV2 (ViT-S/16)³⁹ as CT feature extractor, $FD_{rad} = 384$, and CONCH (ViT-B/16) as pathology feature extractor, $FD_{path} = 512$. For each case, we thus generate N_{c_j} paired radiology and pathology input vectors. Considering all training set cases with corresponding pathology images, and considering all the CT slices with $> 5\%$ tumor area among these cases, an approximate 2025 paired input vectors were generated to train the fusion module for learning pathology-informed CT biomarkers.

Step (b): Fusion: The fusion step enables learning radiology biomarkers that are correlated with pathology features (which we call pathology-informed CT biomarkers). The fusion step consists of two parts: model training and encoding. Inspired by CorrSigNIA,³⁸ the model considered in this work is a Correlational Neural Network (CorrNet).⁵¹ CorrNet learns a common representation between two input vectors, such that they are maximally correlated in this lower dimensional latent space, while at the same time minimizing the reconstruction

error between the output of the decoder and the input. The reconstruction loss prevents collapse of the latent representation (e.g., all zeros).

The CorrNet model consists of one encoder-decoder pair for each input domain (Figure 4). Using one vector pair from the training set, the input vectors R_i and P_i are constructed, originating from radiology and pathology data respectively. The latent representations $h_R = f(WR_i + b)$ and $h_P = f(VP_i + b)$ are the output from the single-layer encoders of each domain, respectively. The activation function f can be any linear or nonlinear function, and $W \in \mathbb{R}^{k \times 384}$, $V \in \mathbb{R}^{k \times 512}$, and $b \in \mathbb{R}^{k \times 1}$, where k is the size of the latent representation. The joint hidden representation is given by

$$h(R_i, P_i) = f(WR_i + VP_i + b). \quad (1)$$

The single layer decoder functionality can be expressed as: $R'_i = g(W'h_R + b')$ and $P'_i = g(V'h_P + b')$, g is any activation function. The loss consists of two parts. The first part, denoted L_{recon} , minimizes the reconstruction error between the input and the output of the CorrNet model. The second part, denoted L_{corr} , maximizes the correlation between the latent representations, h_R and h_P . The total loss is therefore:

$$L = L_{recon} - \lambda L_{corr}, \quad (2)$$

where L_{recon} is given by

$$L_{recon} = \sum_{i=1}^N (L(R_i, R'_i) + L(P_i, P'_i)) \quad (3)$$

and N is the number of samples in one batch and L is the reconstruction error calculated using mean squared error. L_{corr} is given by

$$L_{corr} = \frac{\sum_{i=1}^N (h_R^i - \bar{h}_R)(h_P^i - \bar{h}_P)}{\sqrt{\sum_{i=1}^N (h_R^i - \bar{h}_R)^2 (h_P^i - \bar{h}_P)^2}}. \quad (4)$$

In line with CorrSigNIA,³⁸ we used $\lambda = 2$ for all experiments, the dimension of the hidden space was set to $k = 5$, and f and g are set as the identity function.

Once the model is trained, only the encoder part of the radiology domain is used, and applied to radiology data alone. This ensures the method is usable for clinical inference before surgery or biopsy, where pathology data is unavailable. For inference, the input radiology features per-slice are fed to the trained CorrNet model and the latent

representation is considered the pathology-informed CT biomarkers (denoted as CorrFeat in short). Thus, for a case j with $(N_{c_j}, 384)$ CT feature representations corresponding to N_{c_j} CT slices, the pathology-informed CT biomarkers would have a feature representation of dimension $(N_{c_j}, 5)$. A case-level pathology-informed CT biomarker with five features is then formed by averaging the representations across all the slices.

Step (c): Classification: The classification module to classify ccRCC tumors as aggressive or indolent consists of a multi-layer perceptron (MLP) that is trained with the pathology-informed CT biomarkers and three clinical variables (tumor diameter, gender, age) as inputs. Since the pathology-informed CT biomarkers are learned from resized tumor regions irrespective of the tumor size, the tumor diameter is an orthogonal complementary feature to the imaging biomarkers. Gender and age also provide complementary clinical information to the imaging features. The MLP classifier consists of two fully-connected layers, after the input layer. The first hidden layer after the input layer consists of five nodes and rectified linear unit activation, and the output layer consists of one (1) output node with sigmoid activation function for binary classification. The classifier was trained with binary cross entropy loss function with L2 regularization.

2.3 | Training and test set split

A total of 372 cases were included from KiTS21 and TCGA-KIRC. The training and validation set included 136 indolent cases, and 162 aggressive cases. A test set of 74 cases with balanced distribution of aggressive and indolent ccRCC was randomly selected from KiTS21, to ensure that the test set included cases of high segmentation quality. The remaining cases not included in the test set ($N = 298$) formed a five-fold cross-validation setup. As only TCGA-KIRC included CT and pathology data, this cohort was used to train in steps (a) and (b), while training cases from KiTS21 were included in Step (c). The motivation for using a subset of the KiTS21 dataset in addition to the TCGA-KIRC cohort for training the classification module was to increase the training set size, and also to include heterogeneity in the training set population. In addition, including the KiTS21 dataset to the training set ensured representation of all tumor grades in the training set, as the TCGA-KIRC dataset does not contain any ISUP or Fuhrman Grade 1 clear cell renal cell tumors, but the KiTS21 dataset does.

2.4 | Ablation experiments

The CorrFABR design choices were based on ablation experiments to assess the following:

1. how to optimally extract and aggregate features from whole slide pathology images for registration-independent radiology-pathology correlation learning;
2. how to design an optimum method for aggressiveness classification of ccRCC using CT imaging features;
3. whether pathology-informed CT biomarkers improve classification accuracy;
4. whether classification accuracy can be further improved if complementary clinical information of tumor diameter, gender, and age are included as additional input features.

Evaluation: All models in the ablation experiments were trained and validated using the same dataset ($N = 298$), using the same five-fold cross-validation splits. All models were also evaluated using the same held-out test set ($N = 74$). Since each ablation experiment includes models trained in a five-fold cross-validation setting, the average and standard deviation values of each metric on the test set from the five-fold models are reported. The standard deviation values provide an indication of how stable the model performances are to variations in training and validation splits. In order to compute 95% confidence intervals, and for statistical analysis, the five predicted probability values from the five-fold models for each test-set case were averaged to derive an aggregated per-case predicted probability value. Next, we describe in detail the ablation experiments we performed as part of points 1–4 above, and the corresponding evaluation method for each experiment.

2.4.1 | Feature extraction and aggregation from pathology WSI—The diagnostic pathology WSI were widely different in resolution and image sizes, in comparison to the CT images. To enable radiology-pathology correlated feature learning, two approaches were considered for feature extraction and aggregation from the pathology images: (a) whole image *low-res*, and (b) patch-based *high-res*. In the *low-res* approach, the whole slide pathology images were downsampled and resized to 448×448 pixels, and pathology features were extracted from this severely downsampled pathology images for aggregation. The *low-res* approach follows the prior CorrSigNIA study³⁸ on radiology-pathology correlation learning.

For the *high-res* approach, patches of size 224×224 pixels were extracted from the high-resolution version of the WSI, and the mean of the representations from the selected patches were used to form a representative pathology feature representation for each case. Since there were no cancer annotations on the pathology images in the public dataset and the diagnostic pathology WSI would have a mixture of normal and cancerous tissue, two approaches were adopted for patch-selection: (a) selecting all patches with tissue (after segmenting the background using OTSU thresholding), and (b) selecting cancer patches using zero-shot classification via the pathology vision-language foundation model CONCH.⁴⁰

Evaluation: The following evaluation methods were adopted to decide the optimum pathology feature extraction and aggregation method:

1. Whole image *low-res* versus the patch-based *high-res*: To compare the effects of *low-res* versus the patch-based *high-res* methods of pathology feature extraction and aggregation, the final classification performance of clear cell renal cell carcinoma using each of these approaches, was considered. Classification performance was assessed using the Area Under the Receiver Operating Characteristics Curve (ROC-AUC), F1-score, sensitivity, and specificity. The softmax threshold for the latter metrics was set to 0.5.
2. The two patch-based *high-res* versions (all patches vs. CONCH-selected patches): Since the WSI did not have cancer segmentations, an expert pathologist at our institution (E.C.) annotated the cancer regions on the WSI for 20

cases (10 aggressive ccRCC, and 10 indolent ccRCC). The patches from both methods (i.e., all patches with tissue, and CONCH-selected patches) were then quantitatively and qualitatively evaluated relative to the expert pathologist on these 20 cases. For quantitative evaluation of patch-selection, the metric positive predictive value was considered. A high positive predictive value ensures that selected patches have a high probability of containing cancerous tissue, thereby being more informative about ccRCC aggressiveness. The threshold for the CONCH-predicted normalized cancer similarity was chosen such that a high positive predictive value was achieved for the selected patches.

2.4.2 | Designing an optimum method for classification using CT imaging features

The classification performance of a model depends on the kind of features and the kind of classification model being used. To assess what would be the optimum classifier for our task, our study included experiments with (a) different CT feature extractors (DinoV2,³⁹ BiomedCLIP,⁵² and VGG16⁵³), and different pathology feature extractors (CONCH,⁴⁰ and VGG16⁵³), (b) different levels of granularity for feature extraction and aggregation (slice-level, and tumor-texture level), and (c) different classification modules (MLP based classifiers, and segmentation models with classification heads). The MLP based classifiers were trained with slice-level representations, with or without radiology-pathology correlation learning (Figure 5). The segmentation models with classification heads were trained with either CT image slices or correlated tumor-texture feature maps from radiology-pathology correlation learning (Figure 6). The goal for these experiments was also to see if registration-independent radiology-pathology correlation learning would be better suited for a particular kind of features (convolutional neural networks [CNN]-based or vision transformer-based), or a particular level of representation granularity (slice-level or tumor texture-level). Below, we summarize the experiments for different classification models and the segmentation models with classification heads, respectively.

- Classification Models:

The classification models were transfer learning-based models, using either CT image slices directly as inputs, or features from pre-trained foundation models as inputs. For the transfer learning-based model with CT image slices as input, we fine-tuned the VGG16 classification model (pretrained with ImageNet) by refining the last layer only to avoid overfitting (Model Name: VGG16-fine-tune). For classification models with foundation model-derived features (Figure 5), each model had three components:

- CT feature extractors: Slice-level features were extracted for all the slices in the CT volume using the following pre-trained foundation models:

1. Vision transformer-based DinoV2³⁹ trained on natural images (extracted feature dimension 384 per slice)
2. Vision transformer-based BiomedCLIP⁵² from the vision-language foundation model trained on biomedical image-text pairs (extracted feature dimension 512 per slice)

– Feature Selectors: Since the extracted feature dimensions (384 or 512) were much larger than the training set size (238), feature selection was used to select the most representative features, while reducing the feature dimensionality.

1. In absence of radiology-pathology correlation learning, the linear PCA or the nonlinear Kernel PCA was used to select the most discriminative features.
2. When radiology-pathology correlation learning was used, CorrNet⁵¹ models were trained to learn the optimum combination of radiology features correlated with pathology features (i.e., learning pathology-informed CT biomarkers - CorrFeat). Pathology feature extraction and aggregation was done using CONCH as described in Section 2.2.2 (Step b).

– Classifier: The optimum classifier was decided-based on experiments with several configurations of MLP models with different number of layers, different number of nodes in each layer, and different activation functions. Our experiments showed that the classifier described in Section 2.2.2 (Step c). was the optimum, with more complex models leading to overfitting.

The models are named based on the feature extractor and selector, as summarized in the table in panel B of Figure 5. More details can be found in Section I (Ablation Studies) and Tables I and II of the supplementary material.

- **Segmentation models with classification heads:**

The motivation of segmentation models with classification heads were two-fold: (a) enable end-to-end training with information within the tumor pixels only, excluding pixels outside the tumor mask (in contrast with slice-level information for the MLP classifiers), and (b) investigate the performance of registration-independent radiology-pathology correlation learning on low-level tumor texture features for this task, in line with the prior CorrSiGNIA study for the prostate. Figure 6 summarizes the segmentation models with classification heads for our task.

- In the absence of radiology-pathology correlation learning, a Holistically Nested Edge Detection (HED) segmentation model⁵⁴ was trained end-to-end with three adjacent CT slices to capture the 3D tumor information, with the label being the entire tumor annotated as aggressive or indolent. The predicted probability maps for the 3D tumor volume were then used to derive a binary classification label either via majority voting, or via training an additional MLP classifier with the probability percentiles. (Model names: HED-Majority or HED-XX, where XX is the percentile of the probability maps used for classification)
- When radiology-pathology correlation learning was used, pathology-informed CT biomarkers learned with low-level tumor textures were used

as inputs to the HED segmentation model. For learning correlated features using tumor texture, the second layer of VGG16 was used both as CT and pathology feature extractors, similar to CorrSigNIA (Model Names: VGG16-CorrFeat-HED-Majority or VGG16-CorrFeat-HED-XX, where XX is the percentile of the probability maps used for classification)

Section I (Ablation Studies) and Tables I and II of Supplementary Materials provides a detailed description of the segmentation models with classification heads with and without radiology-pathology correlation learning.

Evaluation: Each method was evaluated using the classification performance in classifying clear cell renal cell carcinoma into aggressive and indolent classes. Classification performance was assessed using the ROC-AUC, F1-score, sensitivity, and specificity. The softmax threshold for the latter metrics were set to 0.5.

2.4.3 | Pathology-informed CT biomarkers and clinical variables as classifier inputs

The motivation for resizing the CT tumor regions for all cases to $224 \times 224 \times N_{c_j}$, (N_{c_j} being the number of CT slices for the case j), irrespective of the tumor size, was to understand the utility of CT imaging features (with or without correlated feature learning) in the classification task. Since tumor diameter can also contain information regarding the aggressiveness of a ccRCC tumor, experiments were done to assess the predictive value of tumor diameter alone and as an additional input feature along with the pathology-informed CT biomarkers. Moreover, motivated by prior studies that indicate that gender and age can be predictors of RCC aggressiveness, we also integrate these two additional complementary clinical variables.

Evaluation: Final classification performance using ROC-AUC, F1-score, sensitivity, and specificity was used to assess the predictive value of tumor diameter alone versus in combination with pathology-informed CT biomarkers.

2.5 | Statistical Analysis

The following statistical tests were performed:

1. Delong's test⁵⁵ was used to compare the ROC-AUCs of our proposed CorrFABR method with the baseline classifier with tumor diameter (since clinicians currently consider tumor diameter as an important predictor), and (b) the best-performing classification model using CT images alone, without any correlated feature learning. Bonferroni correction⁵⁶ was applied for multiple comparisons and the p -value for statistical significance was obtained by dividing by the number of comparisons (i.e., our proposed method would be statistically significantly better if $p < 0.025$).
2. Non-inferiority test was applied to compare ROC-AUCs between methods with the non-inferiority margin of 0.05. If there were multiple comparisons, Bonferroni correction was applied and the threshold of significance was changed correspondingly based on the number of comparisons. Given the small test-set size, and the number of different ablation experiments with incremental

improvements in each step, we chose non-inferiority tests as opposed to Delong's test for the ablation experiments.

3. Two independent power analyses were done. First, power analysis was performed to determine the sample size required to have a certain probability of observing statistically significant differences between CorrFABR with the best-performing classification model with CT images alone, without correlated features or clinical variables in item 1(b). Second, power analysis was also done to calculate the probability of finding a statistically significant difference between the aggressive and indolent classes using our proposed CorrFABR method with our given test set size. For each power analysis, the effect size was computed by dividing the difference between the average predicted probability values of the two groups by the common error standard deviation.

The statistical analysis was done in R version 4.4.1.

2.6 | Training details

Five-fold cross-validation was used to train both the CorrNet model and the classification models. The CorrNet models were trained for 500 epochs, with a learning rate $0.5 \cdot 10^{-4}$, batch size 50, and RMSprop optimizer. The MLP classification model was trained for maximally 5000 epochs with early stopping, a learning rate of 0.001, batch size of 32, and an Adam optimizer. Early stopping was used with patience of 50 epochs, and the model with the best validation performance was used for evaluation.

The HED segmentation models were trained for maximally 100 epochs, a learning rate of 0.001, and batch size of 8. The learning rate was reduced by 0.1 if the validation loss had not been reduced in 10 epochs. Early stopping was used with patience of 20 epochs, and the final model was used for evaluation.

3 | EXPERIMENTAL RESULTS

Summary:

Our experiments demonstrate that simple MLP classifiers that use (a) pathology-informed CT biomarkers learned through registration-independent radiology-pathology correlation learning from foundation model-derived features, and (b) orthogonal clinical variables of tumor diameter, gender, and age as inputs, achieve the optimum performance in classifying ccRCC into aggressive and indolent classes. We also find that the proposed registration-independent radiology-pathology correlation learning helps improve classification performance, irrespective of the feature extractor (CNN-based or vision-transformer-based) and the backbone classification model (classification vs. segmentation with classification head models). Moreover, patch-based *high-res* version of pathology feature extraction and aggregation shows improvement over the *low-res* pathology feature extraction. In the following subsections, we present the results from each of the ablation experiments detailed in Section 2.4, that justify the design choices for CorrFABR.

3.1 | Results of ablation experiments

3.1.1 | Feature extraction and aggregation from pathology WSI

- A. Whole image *low-res* versus patch-based *high-res* approaches: *High-res* patch-based approach (with pathology patches selected by CONCH) outperforms the *low-res* whole image approach (Table 4) for both DinoV2- and BiomedCLIP-based feature extractors. While the Delong test did not show statistically superior ROC-AUC due to the small test set, non-inferiority tests with Bonferroni correction demonstrate statistical significance (both $p < 1e^{-10}$). Standard deviation values in the *high-res* approach for both DinoV2- and BiomedCLIP-based models are lower, suggesting more stable performance to variations in training and validation splits. Higher performance using the patch-based *high-res* approach is expected as information is lost in the severely downsampled *low-res* pathology images. Moreover, for the *low-res* whole slide pathology image feature extraction, the same pathology feature is matched with all the CT slices containing tumor, resulting in less heterogeneity in radiology-pathology correlation learning.
- B. All patches with tissue versus CONCH-selected patches for the *high-res* approach: For the *high-res* patch-based approach for pathology feature extraction, when patches with CONCH-predicted normalized cancer similarity score ≥ 0.9 were selected, an average positive predictive value (PPV) of 0.76 was achieved. This suggests that approximately 76% of CONCH-selected patches (with normalized cancer similarity score ≥ 0.9) would contain cancer. This PPV was higher than when all patches with tissue were considered (PPV = 0.67), as well as when a normalized cancer similarity score of 0.5 was used to select patches using CONCH (PPV = 0.74). Qualitative evaluation shows that CONCH patches with normalized cancer similarity score ≥ 0.9 shows very good overlap with an expert pathologist outline (Figure 7), leading us to use this approach for CorrFABR.

3.1.2 | CT imaging features versus pathology-informed CT biomarkers for classification

Classification models: Pre-trained slice-level feature representations from the foundation models (DinoV2, BiomedCLIP) followed by dimensionality reduction with principal component analysis (PCA) have poor performance, compared to fine-tuning only the last layer of VGG16 (ROC-AUC of DinoV2-PCA vs. BiomedCLIP-PCA vs. VGG16-fine tune: 0.325 ± 0.130 vs. 0.391 ± 0.114 vs. 0.778 ± 0.024) (Table 5). Since the feature dimensions from pre-trained foundation models (384 or 512) are large compared to the training set size (238), directly training the MLP classifiers without feature selection/dimensionality reduction leads to severe overfitting that was observed in the training and validation loss curves. When Kernel PCA (KPCA) with a nonlinear radial basis function kernel was used for feature selection, both DinoV2 and BiomedCLIP features achieved improved performance (ROC-AUC of DinoV2-KPCA: 0.782 ± 0.026 , BiomedCLIP-KPCA: 0.625 ± 0.117).

Radiology-pathology correlation learning as a feature selector led to improved classification ROC-AUC for both DinoV2 and BiomedCLIP-feature extractors, when compared to PCA or KPCA being used as feature selectors (ROC-AUC of Dinov2-CorrFeat-Avg vs. DinoV2-KPCA: 0.815 ± 0.007 vs. 0.782 ± 0.026 ; BiomedCLIP-CorrFeat-Avg vs. BiomedCLIP-KPCA: 0.794 ± 0.008 vs. 0.625 ± 0.117). While the Delong's test did not show statistically superior AUC on account of the small test set, the non-inferiority tests show statistical significance for these comparisons (both $p < 0.016$). The standard deviation values also decrease, suggesting more stable performance to variations in training-validation splits. The improvement in ROC-AUC is larger for BiomedCLIP than for DinoV2, which may be attributed to the difference in size and diversity of the training datasets of these foundation models (BiomedCLIP was trained with 15 million biomedical image-text pairs, whereas DinoV2 was trained with a much larger and diverse dataset of 142 million images through self-supervision). Our experiments suggest that for cancer-specific classification tasks, radiology-pathology correlation learning enables learning the combination of radiology features that are correlated with pathology features of cancer (i.e, pathology-informed radiology biomarkers), thereby providing a means of discriminative feature selection from pre-trained foundation model-derived features. Averaging the slice-level pathology-informed CT biomarkers per-case leads to a better classification accuracy than considering both the average and standard deviation of the slice-level biomarkers on a per-case basis for both the foundation model-derived features.

Segmentation models with classification heads: Similar trend of improved performance with pathology-informed CT biomarkers was also observed with the segmentation models with classification heads (Table 5). It may be noted that for these models, radiology-pathology correlation learning was performed on a tumor texture level, rather than on a slice-level, unlike the classification models. Also, the segmentation models were trained end-to-end without any pre-training. Among the segmentation models with classification heads, VGGCorrFeat-HED-90 model exhibited the best overall performance with an ROC-AUC of 0.767 ± 0.030 , outperforming its counterpart with input CT slices (HED-90, ROC-AUC: 0.729 ± 0.025 , non-inferiority test $p < 0.016$). More experiments with different design choices for the segmentation models with classification heads are reported in Section II (Results) and Table III of the Supplementary Material.

For the models that use majority voting from the predicted segmentation labels for classification, presenting an ROC-AUC number is not straight-forward and may vary depending on how the segmentation probability maps are converted to a probability value for ROC-AUC computation. When the segmentation model-predicted probability maps were used to additionally train MLP classifiers, presenting ROC-AUC numbers are straight-forward as it is based on the predicted probability value from the MLP classifiers. However, not much difference in performance was noted between VGGCorrFeat-HED-XX models and the VGGCorrFeat-HED-Majority model, with the models using the 75th or the 90th percentiles from the predicted probability maps yielding the best results.

Overall, we note that when registration-independent radiology-pathology correlation learning is used, the learned pathology-informed CT biomarkers improve classification performance (Figure 8), irrespective of (a) the kind of CT feature extractors (DinoV2,

BiomedCLIP, or VGG16), (b) the level of feature granularity (slice-level or tumor-texture level), and (c) the used models (classification models or segmentation models with classifier heads). The improvement in performance is larger if the baseline features are from models initially trained with smaller and less diverse datasets (e.g., BiomedCLIP or VGG16), than when the baseline features are from models trained with larger and more diverse datasets (e.g., DinoV2).

3.1.3 | Pathology-informed CT biomarkers + clinical variables for

classification—Tumor diameter alone when used as an input in a Random Forest classifier showed a classification ROC-AUC of 0.679 ± 0.023 . Adding tumor diameter, gender, and age showed improvements in classification performance, irrespective of the type of feature extractors and classifiers (Table 6). Our proposed CorrFABR method that uses pathology-informed CT biomarkers with clinical variables as inputs to an MLP-classifier shows the best performance with an ROC-AUC of 0.855 ± 0.005 (95% CI: 0.775, 0.947). CorrFABR is statistically significantly better than the baseline models that use tumor diameter ($p < 0.008$). Details of performance when each independent clinical variable is added as input features are presented in Table IV of Supplementary Material.

Since CorrFABR leverages foundation model-derived features and shallow CorrNet for feature selection, it does not involve intensive training and large training datasets. However, larger test sets are required to establish statistically significantly improved performance over the second-best model (see Section 3.2 below).

3.2 | Results of statistical analysis

Our statistical analyses demonstrate the following:

1. CorrFABR is statistically significantly better than the classifier that uses tumor diameter, the current clinical non-invasive feature for assessment (DeLong's test, $p < 0.008$).
2. Our test set size ($N = 74$) is insufficient to show statistically significant improvements of CorrFABR over the best classification model with CT images (DinoV2-KPCA). We need a test set of at least 391 samples to show statistically significant differences between the two models (demonstrated by power analysis with one-tailed two-sample t -test with a computed effect size of 0.232, power of 0.9, significance level 0.05). However, CorrFABR is statistically significantly non-inferior to DinoV2-KPCA (non-inferiority test, significance-level 0.05, $p < 0.01$).
3. Although we do not have enough samples to show statistically significant differences between CorrFABR with the best-performing classifier with CT images alone (DinoV2-KPCA) using Delong's test, we do have enough samples to see a statistically significant difference between the CorrFABR-predicted probability values of the aggressive and indolent cases (power analysis using one-sided t -test, each subgroup size of 37 samples, computed effect size of 1.23, significance level of 0.05 gives a power of 0.999).

4. *High-res* pathology feature extraction and aggregation is statistically significantly non-inferior to *low-res* pathology feature extraction and aggregation (non-inferiority tests with Bonferroni multiple comparison correction, comparing DinoV2-Corrfeat-Avg-Low-res with DinoV2-Corrfeat-Avg, and BiomedCLIP-Corrfeat-Avg-Low-res with BiomedCLIP-Corrfeat-Avg, both $p < 1e^{-10}$).
5. Radiology-pathology correlated feature learning is statistically significantly non-inferior to counterparts without correlation learning (non-inferiority tests with Bonferroni multiple comparison correction, comparing DinoV2-KPCA with DinoV2-Corrfeat-Avg, BiomedCLIP-KPCA with BiomedCLIP-Corrfeat, and HED-90 with VGGCorrfeat-HED-90, all $p < 0.016$).

3.3 | Failure analysis

Table 7 shows the performance of the proposed CorrFABR method on subgroups in the test data, grouped by gender, tumor diameter, and age. The diameter and age values have been binned in four bins with approximately equal number of samples in each bin. Looking at the ROC-AUCs of these subgroups, we see a difference in performance between genders (male vs. female ROC-AUC: 0.799 ± 0.012 vs. 0.895 ± 0.015) and age groups (largest difference between ages ≤ 50 years vs. ages ≥ 71 years: ROC-AUC: 0.788 ± 0.046 vs. 0.927 ± 0.022). The largest performance difference was, however, observed between small tumors (< 2.8 cm diameter) and large tumors (≥ 6.1 cm) (ROC-AUC: 0.903 ± 0.040 vs. 0.419 ± 0.084). A closer examination of the misclassified cases showed that CorrFABR consistently misclassified 15 cases (out of 74 test cases) in at least three out of the five trained models in the five-fold setup. Five of these were indolent cases, but incorrectly classified as aggressive. The remaining 10 cases were aggressive, but incorrectly classified as indolent. The misclassified indolent cases had a significantly larger tumor diameter than the correctly classified ones, and the misclassified aggressive cases had a significantly smaller tumor diameter than the correctly classified ones (indolent: correct 3.28 cm vs. incorrect 6.50 cm, aggressive: correct 7.74 cm vs. incorrect 3.38 cm, both $p < 0.05$, two-sided t -test). Figure 9 shows one example false negative and one example false positive case respectively. As seen in literature,^{12,16} there is a strong correlation between tumor diameter and aggressiveness, where larger tumors tend to be aggressive, and vice versa. This is reflected in the training data and consequently in the models' predictions. Looking at the class distribution of tumors smaller than 4.3 cm in the training data, we see that 34% were aggressive. To further improve performance on small tumors, data collection is needed to balance the classes for this clinically relevant group.

4 | DISCUSSION

We presented a novel, clinically-relevant, multimodal fusion method, CorrFABR, that enables integrating information from pathology images (as pathology-informed radiology biomarkers) and clinical variables for radiology-based classification of ccRCC as aggressive or indolent. CorrFABR does not require spatially aligned radiology and pathology images for learning pathology-informed radiology biomarkers. Moreover, pathology images are not required during inference, making CorrFABR a clinically relevant solution for radiology-based assessment in new patients before surgery or biopsy. Such radiology imaging-

based aggressiveness assessment of ccRCC can help guide treatment plans and prevent unnecessary biopsies or surgeries in patients with indolent disease and/or with other medical conditions that put them at higher risk of adverse effects from invasive procedures. This non-invasive assessment is often particularly important for treatment planning in higher age-group patients, or for small tumors with diameter ≤ 4.5 cm. CorrFABR demonstrates excellent performance in patients over 71 years age, as well as for small tumors with diameter ≤ 4.5 cm, suggesting its clinical utility (Table 7). We trained and evaluated CorrFABR on heterogenous public data of ccRCC patients. We found that CorrFABR demonstrated improved classification performance than other methods that use CT images or tumor diameter as inputs.

As the definition of aggressiveness of ccRCC, we used both high tumor grade (grade 3 or 4) and the presence of necrosis, thereby incorporating two important prognostic factors.^{9,15,57} Previous methods have been presented for tumor-grade classification only, without reference to necrosis.^{24–29,31–34} Therefore, numerical comparison of results is challenging, as the target endpoint is different. This is further amplified by evaluation on internal data only and/or the use of different/multiple input data sequences (un-enhanced, varying delay times of enhanced images, or multiple phases).

Our study has a few noteworthy limitations. First, the method is evaluated on ccRCC and does not include other renal cancer subtypes. This is partly because of the sparsity of other RCC examples in the public dataset, and partly because ccRCC is the most prevalent subtype, and radiologic classification of its aggressiveness is an unmet clinical need. In future work, we intend to extend this study for radiology imaging-based aggressiveness classification of all types of kidney tumors. Second, by using public data, CT images taken at different post-contrast phases were included. This introduces noise and heterogeneity in the signal to the model. A more homogenous dataset with exactly the same post-contrast phase could likely improve the models' performance, but potentially at the cost of generalizability across phases. Our dataset highlights the common clinical scenario where there is variability in post-contrast phases across different institutions and even within the same institution. Third, our dataset size is relatively small. Our results demonstrate that with heterogenous publicly available data, CorrFABR shows improved numerical performance over other models, with statistically significant improvements over the tumor diameter-based classifier, and statistically significant non-inferior performance to the best classifier with CT features but without correlated feature learning or clinical variables (DinoV2-KPCA). Power analysis shows that we have enough power (> 0.9) to see differences between indolent and aggressive classes using CorrFABR-predicted probabilities, but unfortunately not enough samples to show significant differences between CorrFABR and the best CT-based classifier. The required sample size to show this statistically significant difference (391) is considerably larger than both our test set and training sets. Future studies will focus on increasing the cohort size, both for training and testing. Fourth, we could not include other clinical variables that may be relevant for this classification, for example, smoking history. This is attributed to incomplete clinical variables available in the public dataset. Finally, two different grading systems were used for the different cohorts, Fuhrman and ISUP. Previous studies⁵⁸ have shown that ISUP is a better prognostic marker than Fuhrman when using all four grades (1–4). However, when grouping the grades into low and high, only 10 cases out

of 279 differed between the grading systems (all 10 cases were Fuhrman grade 3 but ISUP grade 2). The impact of using two grading systems in this study is therefore considered to be low. In fact, a larger variance may come from high inter-reader variability within each tumor grade system.⁵⁷

Our study is the first to bring pathology information into the radiology domain for ccRCC characterization in the clinically relevant scenario when radiology and pathology images are not registered, and pathology images are not needed during inference. Our study is also the first study to leverage foundation model-derived features from both radiology and pathology images to learn pathology-informed radiology biomarkers. Our experiments demonstrate, even with foundation model-derived features, there remains a need to identify relevant radiology cancer biomarkers for downstream medical imaging tasks. This need is stronger when the foundation models are trained with smaller and less heterogeneous datasets, as evidenced by the larger performance improvements with pathology-informed radiology biomarkers in the classifiers that use BiomedCLIP and VGG16 backbones in contrast to classifiers that use DinoV2 backbones. Registration-independent radiology-pathology correlation learning provides a computationally efficient method to identify discriminative and relevant radiology cancer biomarkers from foundation model-derived features, and these learned radiology biomarkers can be further applied to different downstream tasks with simple classifiers, and relatively small datasets. Our promising results also demonstrate the applicability of our approach to other diseases with unaligned radiology and pathology images. Moreover, our study is the first to incorporate clinical variables of tumor diameter, gender, and age in classifying ccRCC aggressiveness. Our experiments demonstrate that these clinical variables provide complementary information that helps improve classification performance. Future work will focus on increasing the cohort size, using different subtypes of kidney cancers, and using larger, multi-institution data to improve classification accuracy and generalizability.

5 | CONCLUSION

In this paper, we present CorrFABR, a method of learning radiological image biomarkers by incorporating pathology image features into the radiology domain without the need for spatially aligned data. CorrFABR also enables integrating orthogonal clinical variables together with pathology-informed radiology biomarkers to distinguish aggressive and indolent ccRCC tumors. During inference, the pathology-informed radiology biomarkers can be extracted without needing pathology images. Our experiments show that CorrFABR shows numerical improvements over all other models, including statistically significant improvement over the tumor diameter-based classifier. CorrFABR also demonstrated statistically significant non-inferior performance to the best CT-based classifier. While pathology-informed radiology biomarkers improve performance irrespective of the pre-trained feature backbone, it was observed that the improvement in performance was larger when the feature extractor backbone has originally been trained with smaller, less diverse datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the Wallenberg AI and Autonomous Systems and Software Program (WASPAI), the research environment ELLIIT, AIDA Vinnova grant 2017-02447, Linköping University Center for Industrial Information Technology (CENIIT), Departments of Radiology and Urology, Stanford University, National Institutes of Health, National Cancer Institute (U01CA196387, to J.D.B., R37CA260346 to M.R.), and the generous philanthropic support of our patients (G.A.S.). Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R37CA260346. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding information

WASP-AI; ELLIIT; AIDA Vinnova grant, Grant/Award Number: 2017-02447; Foundation for the National Institutes of Health, Grant/Award Numbers: U01CA196387, R37CA260346; School of Medicine, Stanford University; Linköping University Center for Industrial Information Technology

DATA AVAILABILITY STATEMENT

The data used in this study is publicly available.

REFERENCES

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–263. <https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21834> [PubMed: 38572751]
2. Gupta K, Miller JD, Li JZ, Russell MW, Charbonneau C. Epidemiologic and socioeconomic burden of metastatic renal cell carcinoma (mRCC): a literature review. *Cancer Treat Rev.* 2008;34(3):193–205. [PubMed: 18313224]
3. Capitanio U, Montorsi F. Renal cancer. *Lancet.* 2016;387(10021):894–906. [PubMed: 26318520]
4. Capitanio U, Bensalah K, Bex A, et al. Epidemiology of renal cell carcinoma. *Eur Urol.* 2019;75(1):74–84. [PubMed: 30243799]
5. de Leon AD, Pedrosa I. Imaging and screening of kidney cancer. *Radiol Clin N.* 2017;55(6):1235–1250.
6. Hsieh JJ, Purdue MP, Signoretti S, et al. Renal cell carcinoma. *Nat Rev Dis Primers.* 2017;3:17009. [PubMed: 28276433]
7. Rasmussen R, Sanford T, Parwani AV, Pedrosa I. Artificial intelligence in kidney cancer. *Am Soc Clin Oncol Educ Book.* 2022;42:1–11.
8. Padala SA, Barsouk A, Thandra KC, et al. Epidemiology of renal cell carcinoma. *World J Oncol.* 2020;11(3):79–87. [PubMed: 32494314]
9. Minardi D, Lucarini G, Mazzucchelli R, et al. Prognostic role of Fuhrman grade and vascular endothelial growth factor in pT1a clear cell carcinoma in partial nephrectomy specimens. *J Urol.* 2005;174(4, pt 1):1208–1212. [PubMed: 16145371]
10. Zhang L, Zha Z, Qu W, et al. Tumor necrosis as a prognostic variable for the clinical outcome in patients with renal cell carcinoma: a systematic review and meta-analysis. *BMC Cancer.* 2018;18(1):870. [PubMed: 30176824]
11. 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. [PubMed: 26935559]

12. Campbell S, Uzzo RG, Allaf ME, et al. Renal mass and localized renal cancer: AUA guideline. *J Urol.* 2017;198(3):520–529. [PubMed: 28479239]
13. Rosales JC, Haramis G, Moreno J, et al. Active surveillance for renal cortical neoplasms. *J Urol.* 2010;183(5):1698–1702. [PubMed: 20299038]
14. Morshid A, Duran ES, Choi WJ, Duran C. A Concise review of the multimodality imaging features of renal cell carcinoma. *Cureus.* 2021;13(2):e13231.
15. Hötter AM, Karlo CA, Zheng J, et al. Clear cell renal cell carcinoma: associations between CT features and patient survival. *AJR Am J Roentgenol.* 2016;206(5):1023–1030. [PubMed: 26934514]
16. Bhindi B, Thompson RH, Lohse CM, et al. The probability of aggressive versus indolent histology based on renal tumor size: implications for surveillance and treatment. *Eur Urol.* 2018;74(4):489–497. [PubMed: 30017400]
17. Coy H, Hsieh K, Wu W, et al. Deep learning and radiomics: the utility of Google TensorFlow™ inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphase CT. *Abdom Radiol.* 2019;44(6):2009–2020.
18. Sun X-Y, Feng Q-X, Xu X, et al. Radiologic-radiomic machine learning models for differentiation of benign and malignant solid renal masses: comparison with expert-level radiologists. *Am J Roentgenol.* 2020;214(1):W44–W54. [PubMed: 31553660]
19. Tanaka T, Huang Y, Marukawa Y, et al. Differentiation of small (≤ 4 cm) renal masses on multiphase contrast-enhanced CT by deep learning. *Am J Roentgenol.* 2020;214(3):605–612. [PubMed: 31913072]
20. Yan L, Liu Z, Wang G, et al. Angiomyolipoma with minimal fat. *Acad Radiol.* 2015;22(9):1115–1121. [PubMed: 26031228]
21. Han S, Hwang SI, Lee HJ. The classification of renal cancer in 3-phase CT images using a deep learning method. *J Digit Imaging.* 2019;32(4):638–643. [PubMed: 31098732]
22. Zuo T, Zheng Y, He L, et al. Automated classification of papillary renal cell carcinoma and chromophobe renal cell carcinoma based on a small computed tomography imaging dataset using deep learning. *Front Oncol.* 2021;11 : 746750.
23. Uhm K-H, Jung S-W, Choi MH, et al. Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ Precis Oncol.* 2021;5(1):54. [PubMed: 34145374]
24. Ding J, Xing Z, Jiang Z, et al. CT-based radiomic model predicts high grade of clear cell renal cell carcinoma. *Eur J Radiol.* 2018;103:51–56. [PubMed: 29803385]
25. Kocak B, Durmaz ES, Ates E, Kaya OK, Kilickesmez O. Unenhanced CT texture analysis of clear cell renal cell carcinomas: a machine learning—based study for predicting histopathologic nuclear grade. *Am J Roentgenol.* 2019;212(6):W132–W139. [PubMed: 30973779]
26. Bektas CT, Kocak B, Yardimci AH, et al. Clear cell renal cell carcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of Fuhrman nuclear grade. *Eur Radiol.* 2019;29(3):1153–1163. [PubMed: 30167812]
27. Lin F, Cui E-M, Lei Y, Luo L-p. CT-based machine learning model to predict the Fuhrman nuclear grade of clear cell renal cell carcinoma. *Abdom Radiol.* 2019;44(7):2528–2534.
28. Sun X, Liu L, Xu K, et al. Prediction of ISUP grading of clear cell renal cell carcinoma using support vector machine model based on CT images. *Medicine.* 2019;98(14):e15022.
29. Shu J, Wen D, Xi Y, et al. Clear cell renal cell carcinoma: machine learning-based computed tomography radiomics analysis for the prediction of WHO/ISUP grade. *Eur J Radiol.* 2019;121: 108738.
30. Nazari M, Shiri I, Hajianfar G, et al. Noninvasive Fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning. *Radiol Med.* 2020;125(8):754–762. [PubMed: 32193870]
31. Cui E, Li Z, Ma C, et al. Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. *Eur Radiol.* 2020;30(5):2912–2921. [PubMed: 32002635]
32. Lin F, Ma C, Xu J, et al. A CT-based deep learning model for predicting the nuclear grade of clear cell renal cell carcinoma. *Eur J Radiol.* 2020;129:109079.

33. Xv Y, Lv F, Guo H, et al. Machine learning-based CT radiomics approach for predicting WHO/ISUP nuclear grade of clear cell renal cell carcinoma: an exploratory and comparative study. *Insights Imaging*. 2021;12(1):170. [PubMed: 34800179]
34. Demirjian NL, Varghese BA, Cen SY, et al. CT-based radiomics stratification of tumor grade and TNM stage of clear cell renal cell carcinoma. *Eur Radiol*. 2021. <https://pubmed.ncbi.nlm.nih.gov/34757449/>
35. Tse JR, Shen L, Shen J, et al. Prevalence of malignancy and histopathological association of Bosniak classification, version 2019 class III and IV cystic renal masses. *J Urol*. 2021;205(4):1031–1038. [PubMed: 33085925]
36. Ning Z, Pan W, Chen Y, et al. Integrative analysis of cross-modal features for the prognosis prediction of clear cell renal cell carcinoma. *Bioinformatics*. 2020;36(9):2888–2895. [PubMed: 31985775]
37. Bhattacharya I, Seetharaman A, Shao W, et al. CorrSigNet: Learning CORRelated prostate cancer SIGNatures from radiology and pathology images for improved computer aided diagnosis. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science*, Cham. Springer International Publishing; 2020:315–325.
38. Bhattacharya I, Seetharaman A, Kunder C, et al. Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework. *Med Image Anal*. 2022;75: 102288.
39. Oquab M, Darcet T, Moutakanni T, et al. . Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* 2023. <https://arxiv.org/abs/2304.07193>
40. Lu MY, Chen B, Williamson DF, et al. A visual-language foundation model for computational pathology. *Nat Med*. 2024;30:863–874. [PubMed: 38504017]
41. Schiavoni V, Campagna R, Pozzi V, et al. Recent Advances in the Management of clear cell renal cell carcinoma: novel biomarkers and targeted therapies. *Cancers*. 2023;15(12): 3207. [PubMed: 37370817]
42. Heller N, Isensee F, Maier-Hein KH, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Med Image Anal*. 2021;67:101821.
43. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045–1057. [PubMed: 23884657]
44. Akin O, Elnajjar P, Heller M, et al. The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC) (Version 3) [Data set]. The Cancer Imaging Archive. doi:10.7937/K9/TCIA.2016.V6PBVTDR
45. Müller D, Kramer F. MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. *BMC Med Imaging*. 2021;21(1):12. [PubMed: 33461500]
46. Otsu NA threshold selection method from gray level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62–66. <https://ieeexplore.ieee.org/document/4310076>
47. Fuhrman SA, Lasky LC, Limas C. Prognostic significance of morphologic parameters in renal cell carcinoma. *Am J Surg Pathol*. 1982;6(7):655–663. [PubMed: 7180965]
48. Beddy P, Genega EM, Ngo L, et al. Tumor necrosis on magnetic resonance imaging correlates with aggressive histology and disease progression in clear cell renal cell carcinoma. *Clin Genitourin Cancer*. 2014;12(1):55–62. [PubMed: 24145001]
49. Zigeuner R, Hutterer G, Chromecki T, et al. External validation of the Mayo clinic stage, size, grade, and necrosis (SSIGN) score for clear-cell renal cell carcinoma in a single European centre applying routine pathology. *Eur Urol*. 2010;57(1): 102–111. [PubMed: 19062157]
50. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Boston, MA, USA. IEEE; 2009:1107–1110.
51. Chandar S, Khapra MM, Larochelle H, Ravindran B. Correlational neural networks. *Neural Comput*. 2016;28(2):257–285. [PubMed: 26654210]
52. Zhang S, Xu Y, Usuyama N, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* 2023.

53. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. 2015.
54. Xie S, Tu Z. Holistically-nested edge detection. *Int J Comput Vision*. 2017;125(1):3–18.
55. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845. [PubMed: 3203132]
56. Weisstein EW. Bonferroni correction. <https://mathworld.wolfram.com/BonferroniCorrection.html>
57. Rabjerg M, Gerke O, Engvad B, Marcussen N. Comparing world health organization/international society of urological pathology Grading and Fuhrman grading with the prognostic value of nuclear area in patients with renal cell carcinoma. *Uro*. 2021;1(1):2–13.
58. Dagher J, Delahunt B, Rioux-Leclercq N, et al. . Clear cell renal cell carcinoma: validation of world health organization/international society of urological pathology grading. *Histopathology*. 2017;71(6):918–925. [PubMed: 28718911]

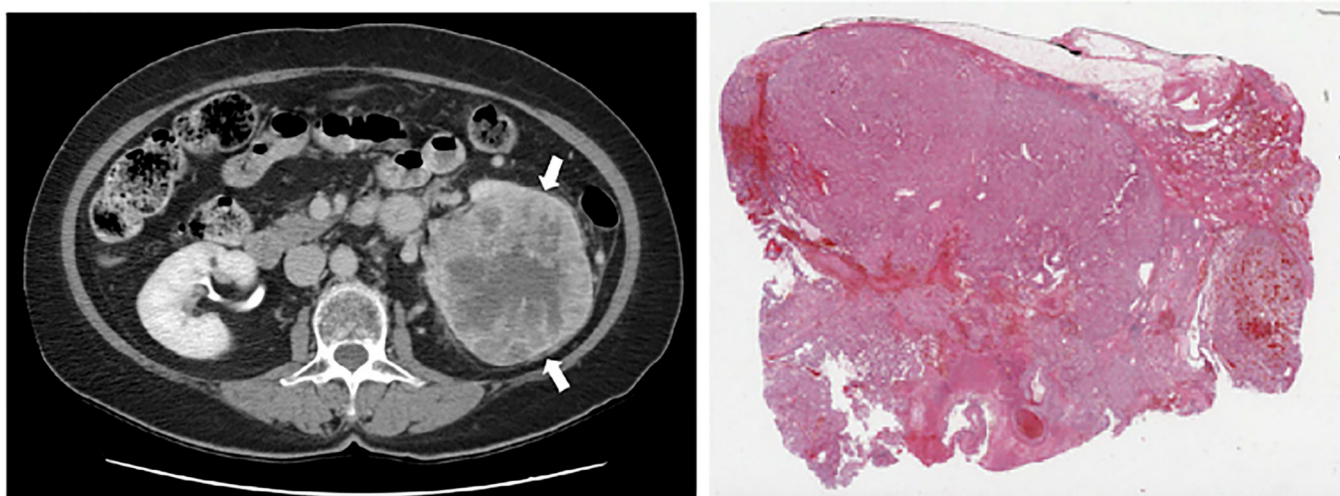
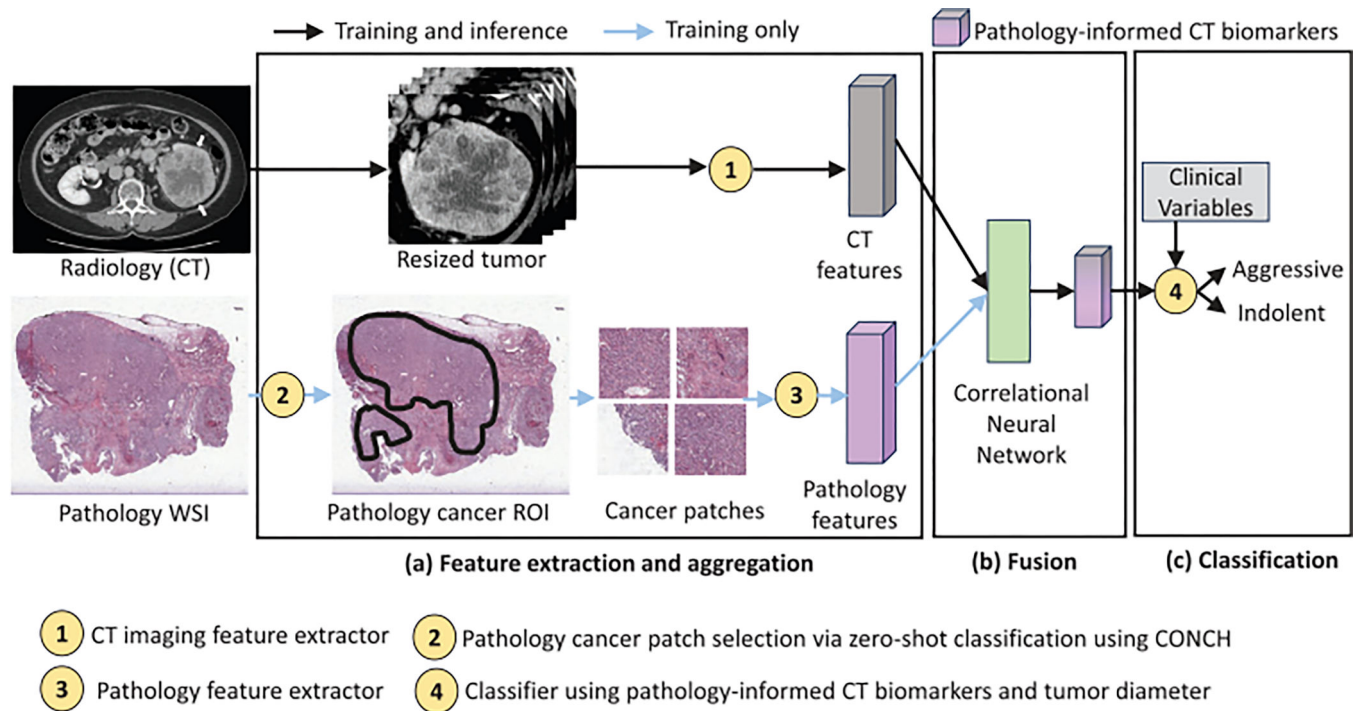
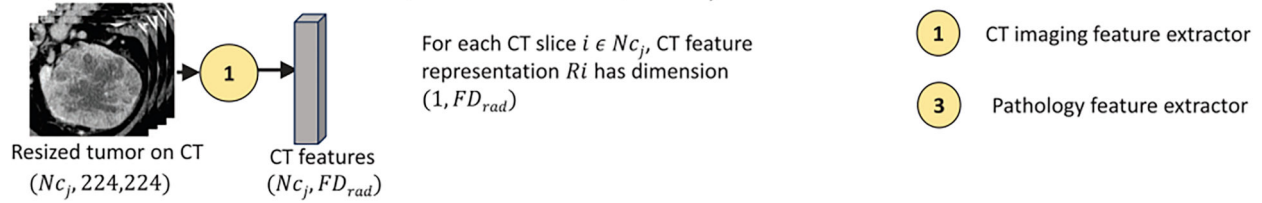
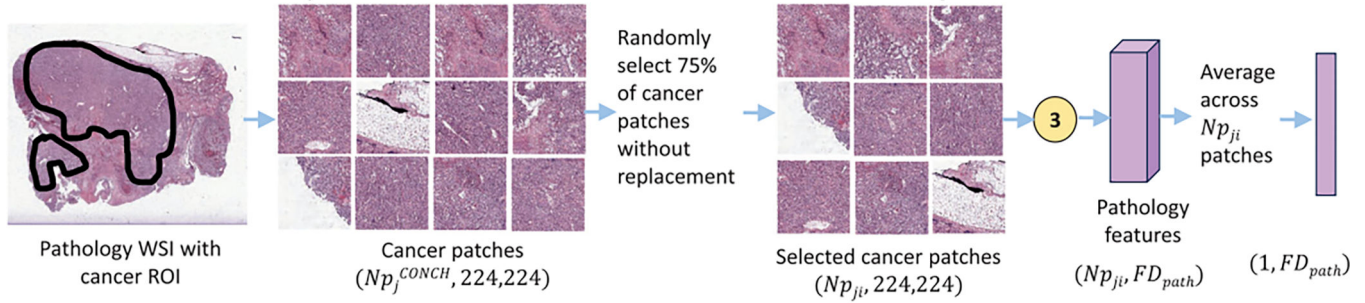


FIGURE 1.

Example images from a patient in our cohort. (Left) CT image with an aggressive tumor highlighted with white arrows, and (Right) the histopathology slice showing surgically resected tissue from an unknown location within the tumor. Please note the differences in appearance and resolution between the CT and histopathology images. CT, computed tomography.

**FIGURE 2.**

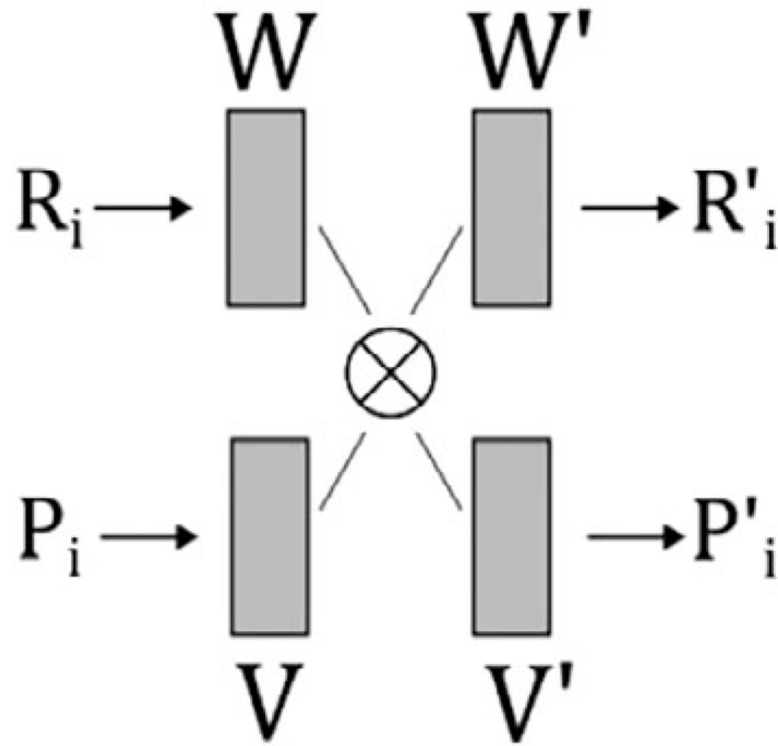
Summary of the proposed method CorrFABR to classify ccRCC into aggressive and indolent classes. The three-step approach consists of (a) a feature extraction and aggregation module to aggregate region-level features from radiology (CT) and pathology images, (b) a fusion module to learn pathology-informed CT biomarkers, and (c) finally, the classification module to classify ccRCC tumors using the learned pathology-informed CT biomarkers and clinical variables (tumor diameter, gender, age). CorrFABR uses the pretrained vision transformer-based foundation model DinoV2³⁹ as the CT feature extractor (circle-1), and the pretrained vision-language foundation model CONCH⁴⁰ to select cancer patches from the WSI, generating a pathology cancer Region-of-Interest (circle-2), as well as to extract pathology features (circle-3). The CorrNet⁵¹ is used to learn pathology-informed CT biomarkers from the extracted and aggregated multimodal features, and a MLP is used as the classifier (circle-4). Pathology images are only required during training in modules (a and b) (light blue arrow), and not required during inference. ccRCC, clear cell RCC; CorrFABR, correlated feature aggregation by region; CorrNet, correlational neural network; CT, computed tomography; ROC-AUC (area under the curve), area under the receiver operating characteristics curve; WSI, whole slide images.

Step A. Generate slice-level CT feature representations for case j with N_{c_j} slices**Step B. For each CT slice $i \in N_{c_j}$, corresponding feature representation P_i with dimension $(1, FD_{path})$ is generated as below:****Step C: For case j with N_{c_j} CT slices, repeat step B for all slices.**

Dimension of case-level CT representation $S_{rad} = (N_{c_j}, FD_{CT})$, and pathology representation $S_{path} = (N_{c_j}, FD_{path})$

FIGURE 3.

Generating pairs of radiology and pathology input features for training the CorrNet model.
CorrNet, correlational neural network.

**FIGURE 4.**

CorrNet model, with R_i and P_i as input representing one pair of feature vectors. The input is processed through the single-layer encoder with weights W and V , respectively. The correlation loss (L_{corr}) forces the resulting representations to be as correlated as possible. The output from the single-layer decoders, W' and V' , reconstructs the input signal as closely as possible through the reconstruction loss (L_{recon}). CorrNet, correlational neural network.

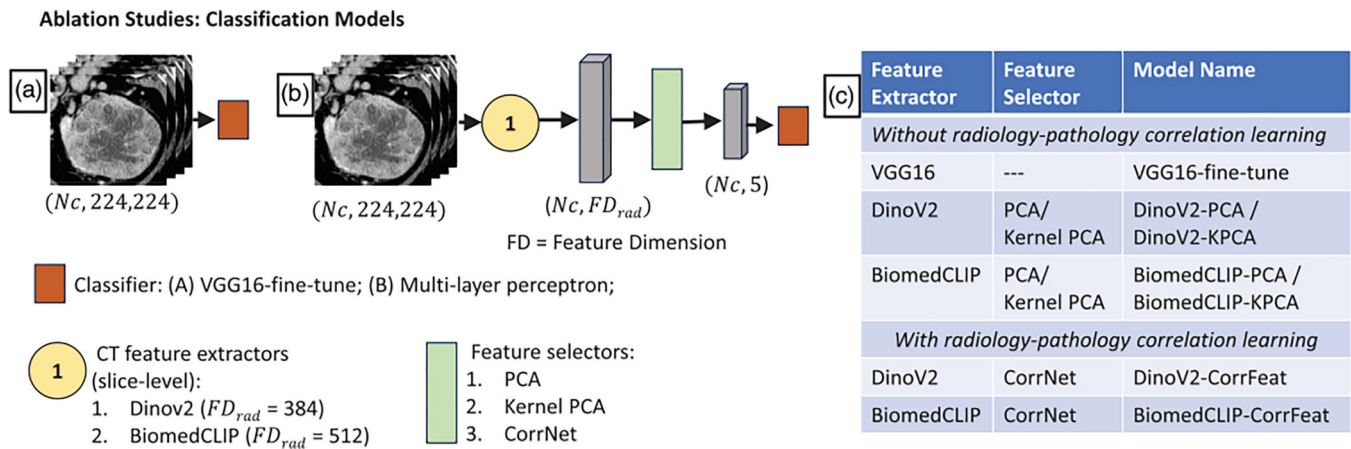


FIGURE 5. Ablation Studies were performed to decide the optimum classification method. (a) The VGG16 classification model was fine-tuned by training the last layer with our dataset, while keeping all other weights from the pre-trained model frozen. (b) For foundation model-derived features, each classification method included three modules: the CT feature extractor (circle-1), the feature selector (green rectangle), and the classifier (brown rectangle). The CT feature extractors used were pretrained DinoV2,³⁹ and BiomedCLIP.⁵² Feature selectors were PCA, kernel PCA and CorrNet⁵¹ for radiology-pathology correlation learning. The classifier was always a MLP. (c) The corresponding model names reflect the kind of feature and feature selector used. CT, computed tomography; CorrNet, correlational neural network; MLP, multi-layer perceptron; PCA, principal component analysis.

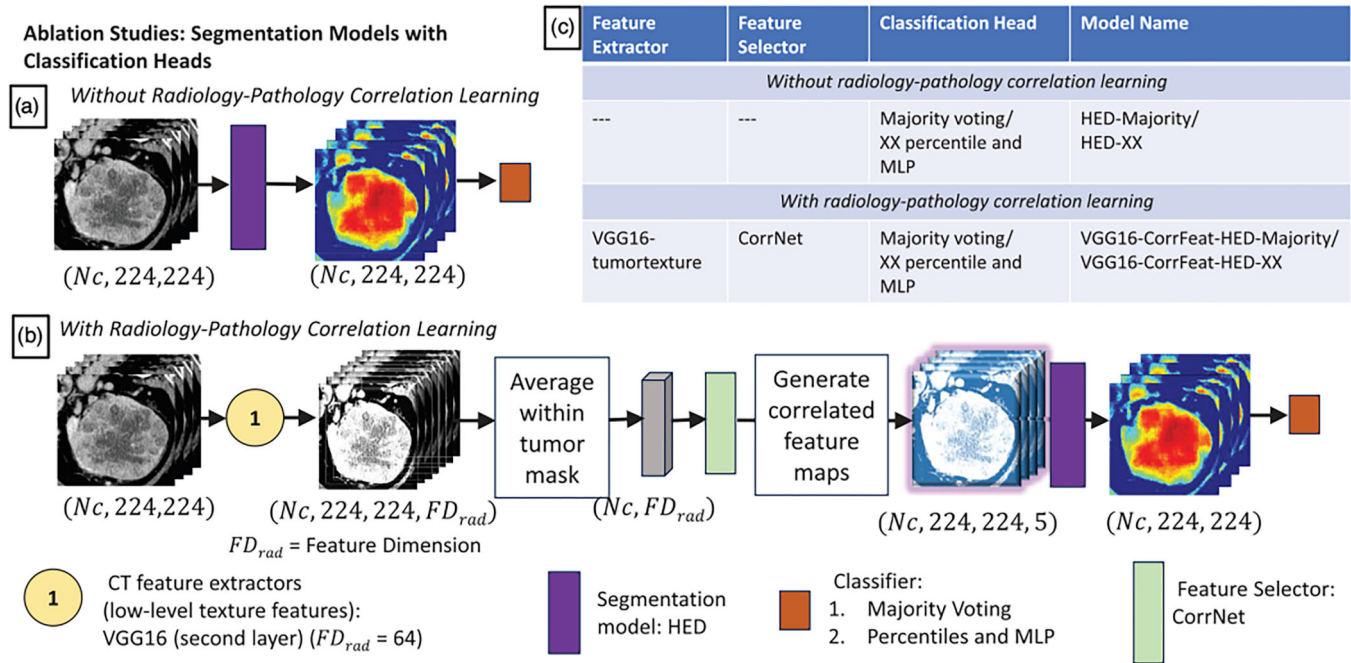


FIGURE 6. Ablation studies were performed with segmentation models with classification heads. (a) Without radiology-pathology correlation learning, input CT images were used as inputs to a segmentation model (purple rectangle), and the predicted probability maps were fed to a classification head (brown rectangle) to convert the segmentation probability maps to binary classification outputs. (b) When radiology-pathology correlation learning was used, low-level texture features were used to learn pathology-informed CT biomarker feature maps for the entire tumor (green rectangle), which were then used as inputs to the segmentation model with classification head. CT, computed tomography.

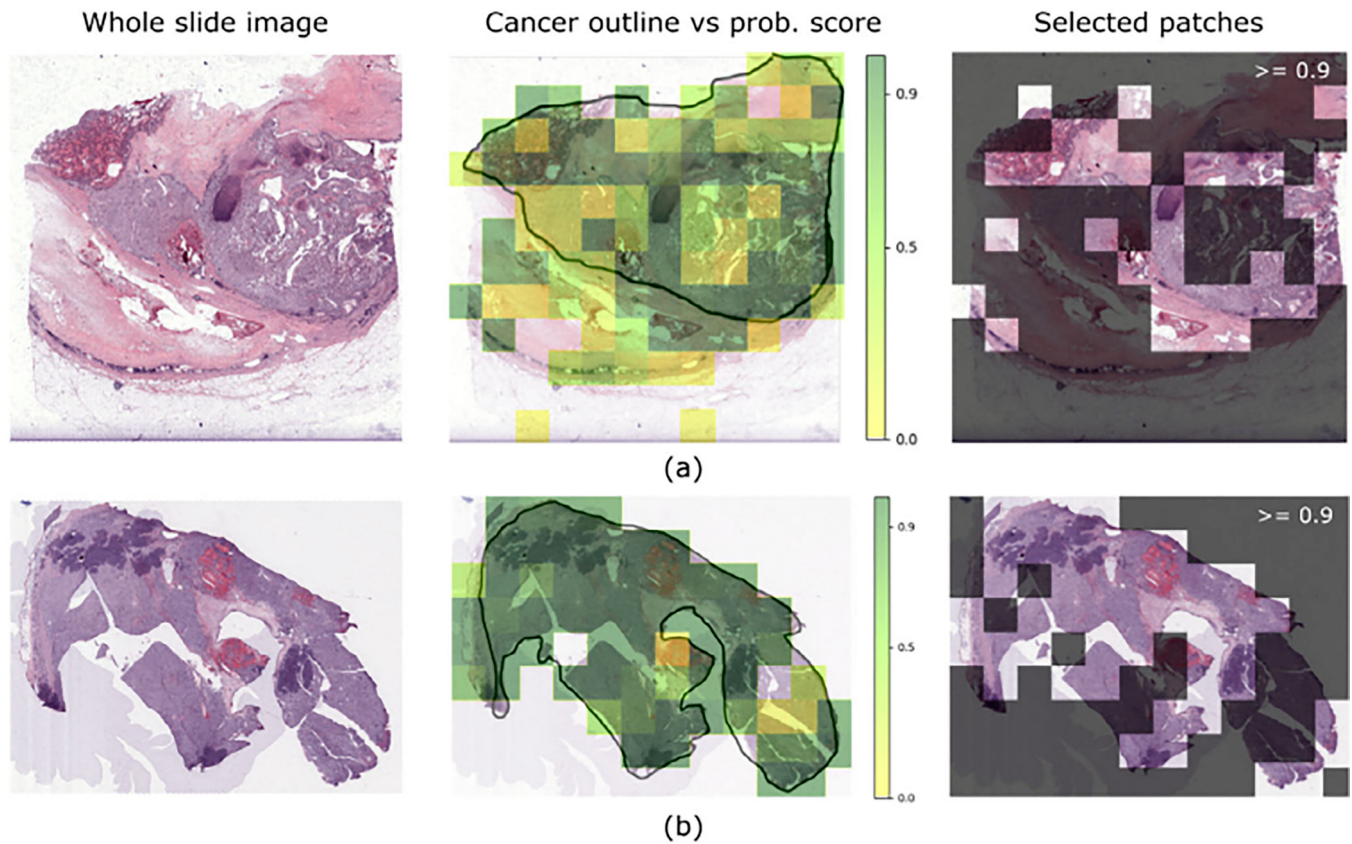
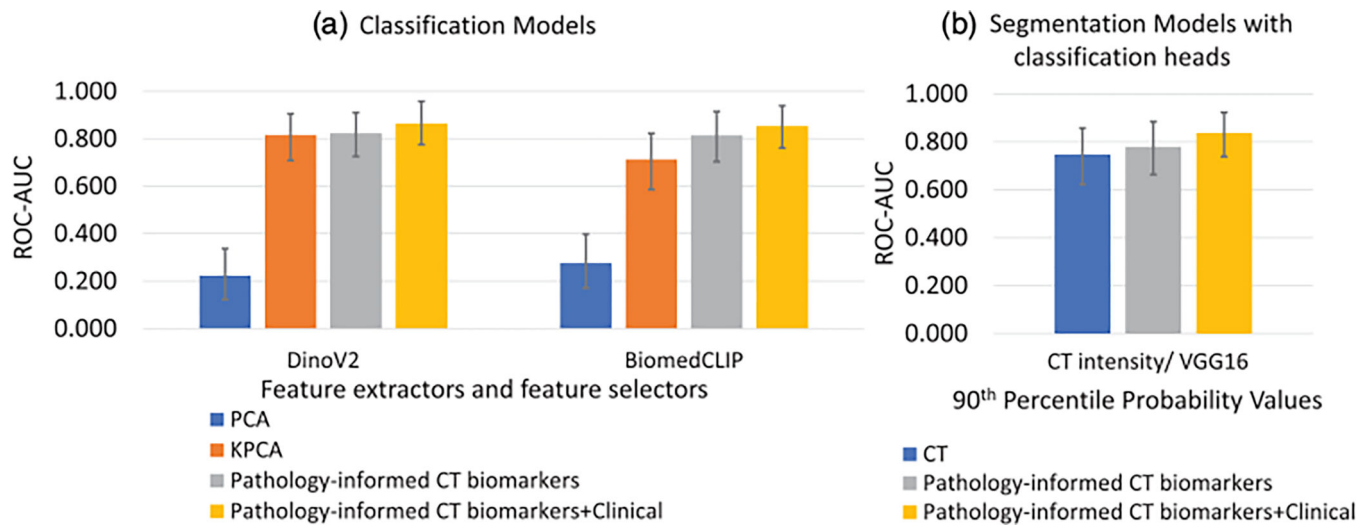
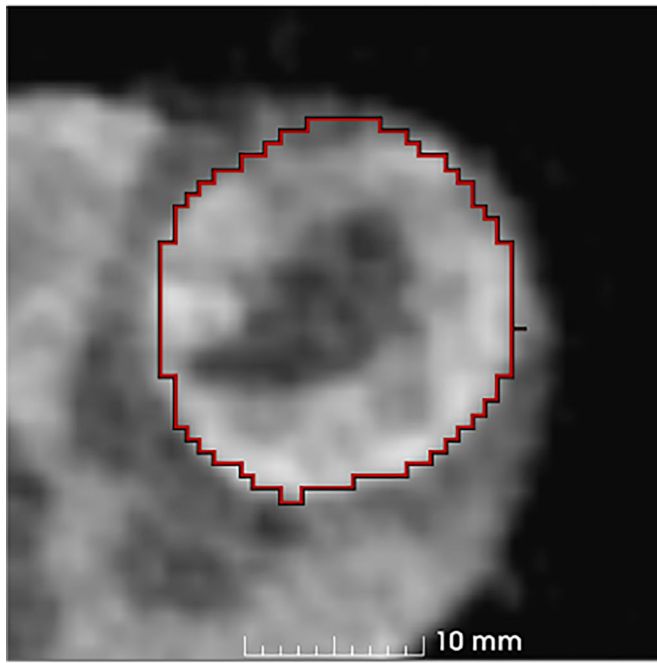


FIGURE 7.

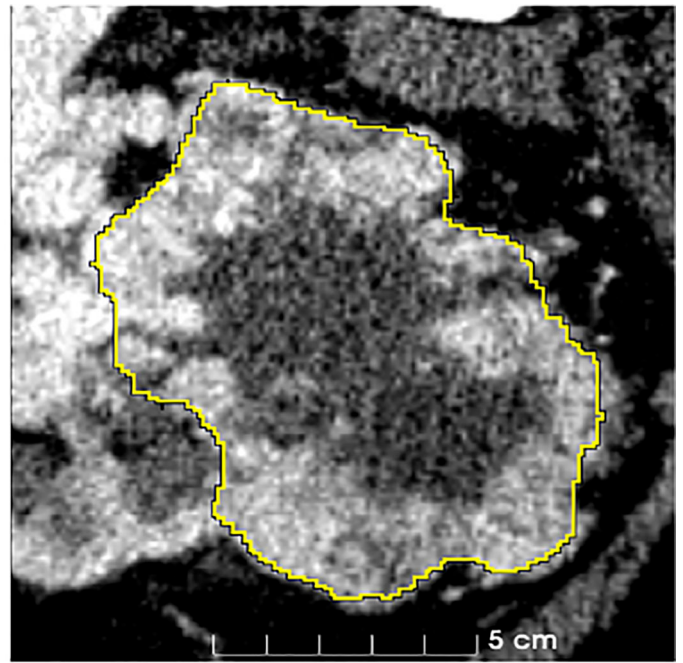
Patch selection using CONCH in a zero-shot classification framework. The center column shows the predicted normalized similarity scores for the tissue patches overlaid on the WSI, with the pathologist's outline of the cancer region in black. Patches were selected using a threshold of CONCH-predicted normalized similarity score ≥ 0.9 . (a) shows an aggressive example, where selected patches (right) and the pathologist outline have a PPV of 0.77 for this particular case. (b) shows an indolent example, where selected patches (right) and the pathologist outline have a PPV of 0.76 for this particular case. Best viewed in color. CONCH, CONTRastive learning from Captions for Histopathology; WSI, whole slide images.

**FIGURE 8.**

Combination of pathology-informed CT biomarkers and clinical variables improve classification performance, irrespective of the kind of CT feature extractors (DinoV2,³⁹ BiomedCLIP,⁵² or VGG16⁵³), the level of granularity for feature extraction and correlation-learning (slice-level or tumor-texture level), and the models (classification models or segmentation models with classifier heads) used. (a) Classification models use slice-level features from DinoV2 or BiomedCLIP, whereas (b) segmentation models with classifier heads use low-level tumor texture features from VGG16. For learning pathology-informed CT biomarkers, all models use CONCH⁴⁰ to select cancer patches from the pathology WSI. Clinical variables (tumor diameter, gender, age) are included as additional orthogonal input features. Each bar represents the ROC-AUC computed from the average of the predicted probabilities from the five-fold models with 95% confidence interval values. CONCH, contrastive learning of histopathology; CT, computed tomography; ROC-AUC, area under the receiver operating characteristics curve; WSI, whole slide images.



(a) False negative: Tumor diameter 2.5 cm, ISUP grade: 3, necrosis present, 59-year-old female



(b) False positive: Tumor diameter 13.2 cm, ISUP grade: 2, necrosis absent, 69-year-old male

FIGURE 9.

One false negative and one false positive example by CorrFABR. (a) The tumor with diameter 2.5 cm was aggressive based on ISUP grade of 3 and presence of necrosis, but CorrFABR incorrectly classifies it as indolent. Tumor segmentation is shown in red outline for this aggressive case. (b) The tumor with diameter 13.2 cm was indolent based on ISUP grade of 2 and absence of necrosis, but CorrFABR incorrectly classifies it as aggressive. Tumor segmentation is shown in yellow outline for this indolent case. CorrFABR, correlated feature aggregation by region; ISUP, international society of urological pathology.

TABLE 1

Dataset summary.

	KITS21	TCGA-KIRC
Number of patients	203	169
Sex (male/female)	133/70	106/63
Age (years)	59.6 ± 12.5 (min: 26, max: 86)	59.5 ± 12.0 (min: 26, max: 88)
Registered	N/A	No
Data	CT	CT
Sequence/Data type	Contrast-enhanced	Contrast-enhanced
Number of slices (per volume)	29–1059	32–306
In-plane resolution (mm)	0.59–1.04	0.50–0.98
Distance between slices (mm)	0.5–5.0	1.25–7.5
Annotation type	Manual	Automatic
		Histopathology
		H and E
		1
		$(0.25 - 0.5) \cdot 10^{-3}$
		N/A
		Not available

Note: Available segmentations of kidney masses were either done manually by trained clinicians or automatically using pre-trained machine-learning methods.
Abbreviation : KITS21, 2021 kidney and kidney tumor segmentation challenge.

TABLE 2

Tumor diameter in cm.

Tumor Grade	Grade 1	Grade 2	Grade 3	Grade 4
KiTS21	3.24 ± 1.17	4.07 ± 2.31	6.27 ± 3.67	8.49 ± 2.70
TCGA-KIRC	—	4.96 ± 2.60	5.82 ± 2.72	8.00 ± 3.01

Note: For KiTS21, radiologic measurements are given. For TCGA-KIRC, approximations of diameter based on calculated volume from automatic segmentations are given. TCGA-KIRC has no Grade 1 cases.

Abbreviation: KiTS21, 2021 kidney and kidney tumor segmentation challenge; TCGA-KIRC, The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection.

TABLE 3

Number of cases with/without necrosis per tumor grade.

	Without necrosis	With necrosis	Total
Grade 1	24	0	24
Grade 2	149	19	168
Grade 3	107	27	134
Grade 4	23	23	46
Total	303	69	372

Note: Bold numbers indicate cases defined as aggressive (with necrosis and/or tumor grade 3 or 4).

Comparing *low-res* and *high-res* approaches of pathology feature extraction and aggregation demonstrates higher performance with the *high-res* approach that uses CONCH-selected patches with a normalized cancer similarity score ≥ 0.9

TABLE 4

Model name	Path. Feat.	five-fold ROC-AUC	95% Confidence interval
<i>Training MLP with pathology informed CT biomarkers</i>			
Dinov2-Corrfeat-Avg-Low-res	<i>Low-res</i>	0.796 $\pm$ 0.018	(0.714, 0.907)
Dinov2-Corrfeat-Avg	<i>High-res</i>	0.815 $\pm$ 0.007	(0.725, 0.910)
BiomedCLIP-Corrfeat-Avg-Low-res	<i>Low-res</i>	0.771 $\pm$ 0.022	(0.672, 0.898)
BiomedCLIP-Corrfeat-Avg	<i>High-res</i>	0.794 $\pm$ 0.007	(0.703, 0.914)

Bold values indicate best performance.

Note: The five-fold ROC-AUC values are reported as mean $\pm$ standard deviation across the five-folds.

Abbreviations: CONCH, Contrastive learning from Captions of Histopathology; MLP, multi-layer perceptron; ROC-AUC, area under the receiver operating characteristics curve.

TABLE 5
Comparing performance of CT imaging features and pathology-informed CT biomarkers in classifying ccRCC tumors.

Model name	ROC-AUC	95% CI	F1-score	Sensitivity	Specificity
Classification Models					
<i>Transfer learning with CT slices or training MLP with foundation-model derived features</i>					
VGG16-fine-tune	0.778 ± 0.024	(0.707, 0.897)	0.654 ± 0.073	0.616 ± 0.143	0.757 ± 0.117
DinoV2-PCA	0.325 ± 0.130	(0.122, 0.336)	0.408 ± 0.087	0.438 ± 0.075	0.27 ± 0.204
DinoV2-KPCA	0.782 ± 0.026	(0.708, 0.905)	0.709 ± 0.022	0.719 ± 0.044	0.692 ± 0.070
BiomedCLIP-PCA	0.391 ± 0.114	(0.170, 0.397)	0.499 ± 0.074	0.573 ± 0.106	0.281 ± 0.122
BiomedCLIP-KPCA	0.625 ± 0.117	(0.585, 0.823)	0.629 ± 0.048	0.692 ± 0.037	0.486 ± 0.112
<i>Training MLP with pathology informed CT biomarkers</i>					
Dinov2-Corrfeat-Avg	0.815 ± 0.007	(0.725, 0.910)	0.663 ± 0.047	0.573 ± 0.083	0.854 ± 0.056
Dinov2-Corrfeat-AvgStd	0.768 ± 0.086	(0.722, 0.911)	0.654 ± 0.086	0.578 ± 0.079	0.811 ± 0.070
BiomedCLIP-Corrfeat-Avg	0.794 ± 0.008	(0.703, 0.914)	0.657 ± 0.028	0.519 ± 0.032	0.941 ± 0.011
BiomedCLIP-Corrfeat-AvgStd	0.765 ± 0.021	(0.693, 0.900)	0.642 ± 0.056	0.541 ± 0.082	0.865 ± 0.057
Segmentation models with classification head					
<i>Training segmentation + classification models with CT images</i>					
HED-75	0.723 ± 0.036	(0.610, 0.849)	0.641 ± 0.027	0.557 ± 0.050	0.822 ± 0.047
HED-90	0.729 ± 0.035	(0.622, 0.857)	0.651 ± 0.024	0.578 ± 0.047	0.805 ± 0.036
<i>Training segmentation+classification models with pathology-informed CT biomarkers</i>					
VGGCorrFeat-HED-75	0.758 ± 0.029	(0.662, 0.881)	0.7 ± 0.053	0.632 ± 0.085	0.832 ± 0.040
VGGCorrFeat-HED-90	0.767 ± 0.030	(0.663, 0.884)	0.694 ± 0.053	0.649 ± 0.089	0.789 ± 0.032

Bold values indicate best performance.

Note: Pathology-informed CT biomarkers improve ROC-AUC irrespective of the kind of features and the models used. DinoV2-CorrFeat-Avg achieves the best ROC-AUC performance, with the lowest standard deviation value across the five-folds.

Abbreviations:ccRCC,clear cell Renal Cell Carcinoma;CT,computed tomography;HED, holistically nested edge detection;KPCA,kernel principal component analysis; MLP, multi-layer perceptron-based; PCA, principal component analysis; ROC-AUC, area under the receiver operating characteristics curve.

TABLE 6
Performance of classifiers with pathology-informed CT biomarkers and complementary clinical variables (tumor diameter, gender, age) as inputs.

Model name	ROC-AUC	95% CI	F1-score	Sensitivity	Specificity
Tumor-Diameter as feature					
Random forest	0.679 ± 0.023	(0.568, 0.815)	0.615 ± 0.032	0.605 ± 0.047	0.638 ± 0.013
Pathology-informed CT biomarkers + clinical variables					
CorrFABR	0.855 ± 0.005	(0.775, 0.947)	0.793 ± 0.029	0.741 ± 0.058	0.876 ± 0.032
BiomedCLIP-CorrFeat-Avg-Clinical	0.841 ± 0.011	(0.761, 0.939)	0.788 ± 0.022	0.768 ± 0.044	0.822 ± 0.037
VGGCorrFeat-HED-90-Clinical	0.802 ± 0.025	(0.737, 0.923)	0.735 ± 0.042	0.730 ± 0.064	0.746 ± 0.056

Bold values indicate best performance.

Note: Our proposed CorrFABR method that uses pathology-informed CT biomarkers and clinical variables in a MLP classifier outperforms other methods. Adding clinical variables improves performance of all models, irrespective of the type of imaging feature extractor (DinoV2, BiomedCLIP, VGG16) and the type of classifier. Abbreviations: CorrFABR, correlated feature aggregation by region; CT, computed tomography; HED, holistically nested edge detection; MLP, multi-layer perceptron-based; ROC-AUC, area under the receiver operating characteristics curve.

TABLE 7

Subgroup performance of the proposed CorrFABR approach.

		# cases (indolent/aggressive)	AUC
Overall		74 (37/37)	0.855 ± 0.005
Gender			
	Male	47 (20/27)	0.799 ± 0.012
	Female	27 (17/10)	0.895 ± 0.015
Diameter (cm)			
	≤ 2.8	18 (12/6)	0.903 ± 0.040
	2.9–4.3	19 (15/4)	0.943 ± 0.015
	4.4–6	19 (8/11)	0.950 ± 0.038
	≥ 6.1	18 (2/16)	0.419 ± 0.084
Age (years)			
	≤ 50	15 (10/5)	0.788 ± 0.046
	51–60	21 (12/9)	0.874 ± 0.024
	61–70	22 (10/12)	0.822 ± 0.013
	≥ 71	16 (5/11)	0.927 ± 0.022

Abbreviation: CorrFABR, correlated feature aggregation by region.