frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fonc.2022.961779

OPEN ACCESS

EDITED BY
Xiongbing Zu,
Xiangya Hospital, Central South University, China

REVIEWED BY
Tiezheng Qi,
Xiangya School of Medicine, Central South University, China
Shicheng Guo,
University of Wisconsin-Madison, United States

*CORRESPONDENCE
Ming Liu
csmliu@uestc.edu.cn
Bin Song
songbin@wchscu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Genitourinary Oncology, a section of the journal Frontiers in Oncology

RECEIVED 05 June 2022
ACCEPTED 29 August 2022
PUBLISHED 28 September 2022

CITATION
Yang M, He X, Xu L, Liu M, Deng J, Cheng X, Wei Y, Li Q, Wan S, Zhang F, Wu L, Wang X, Song B and Liu M (2022) CT-based transformer model for non-invasively predicting the Fuhrman nuclear grade of clear cell renal cell carcinoma. *Front. Oncol.* 12:961779. doi: 10.3389/fonc.2022.961779

CT-based transformer model for non-invasively predicting the Fuhrman nuclear grade of clear cell renal cell carcinoma

Meiyi Yang[1,2†], Xiaopeng He[3†], Lifeng Xu[5], Minghui Liu[2], Jiali Deng[2], Xuan Cheng[2], Yi Wei[4], Qian Li[4], Shang Wan[4], Feng Zhang[5], Lei Wu[2], Xiaomin Wang[2], Bin Song[4*] and Ming Liu[5*]

[1]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, [3]Department of Radiology, Affiliated Hospital of Southwest Medical University, Luzhou, China, [4]Department of Radiology, West China Hospital, Sichuan University, Chengdu, China, [5]Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China

Background: Clear cell Renal Cell Carcinoma (ccRCC) is the most common malignant tumor in the urinary system and the predominant subtype of malignant renal tumors with high mortality. Biopsy is the main examination to determine ccRCC grade, but it can lead to unavoidable complications and sampling bias. Therefore, non-invasive technology (e.g., CT examination) for ccRCC grading is attracting more and more attention. However, noise labels on CT images containing multiple grades but only one label make prediction difficult. However, noise labels exist in CT images, which contain multiple grades but only one label, making prediction difficult.

Aim: We proposed a Transformer-based deep learning algorithm with CT images to improve the diagnostic accuracy of grading prediction and to improve the diagnostic accuracy of ccRCC grading.

Methods: We integrate different training models to improve robustness and predict Fuhrman nuclear grade. Then, we conducted experiments on a collected ccRCC dataset containing 759 patients and used average classification accuracy, sensitivity, specificity, and AreaUnderCurve as indicators to evaluate the quality of research. In the comparative experiments, we further performed various current deep learning algorithms to show the advantages of the proposed method. We collected patients with pathologically proven ccRCC diagnosed from April 2010 to December 2018 as the training and internal test dataset, containing 759 patients. We propose a transformer-based network architecture that efficiently employs convolutional neural networks (CNNs) and self-attention mechanisms to extract a persuasive feature automatically. And then, a nonlinear classifier is applied to classify. We integrate different training models to improve the accuracy and robustness of the model. The average classification accuracy, sensitivity, specificity, and area under curve are used as indicators to evaluate the quality of a model.

Results: The mean accuracy, sensitivity, specificity, and Area Under Curve achieved by CNN were 82.3%, 89.4%, 83.2%, and 85.7%, respectively. In contrast, the proposed Transformer-based model obtains a mean accuracy of 87.1% with a sensitivity of 91.3%, a specificity of 85.3%, and an Area Under Curve (AUC) of 90.3%. The integrated model acquires a better performance (86.5% ACC and an AUC of 91.2%).

Conclusion: A transformer-based network performs better than traditional deep learning algorithms in terms of the accuracy of ccRCC prediction. Meanwhile, the transformer has a certain advantage in dealing with noise labels existing in CT images of ccRCC. This method is promising to be applied to other medical tasks (e.g., the grade of neurogliomas and meningiomas).

KEYWORDS

tumor grading, ensemble learning, clear cell renal cell carcinoma, transformer network, deep learning

1 Introduction

Renal cell carcinoma (RCC) is the most common kidney tumor (1, 2). Clear cell RCC (ccRCC) is the predominant hypotype of RCC, accounting for about 75-80 (3). With the background of population aging, kidney cancer, especially RCC, keeps rising. When ccRCC progresses to intermediate and advanced stages, lymph node metastasis or distant organ metastasis probably occurs, which leads to dangerous clinical symptoms and a poor prognosis (4, 5). At present, one of the most important pieces of clinical evidence for judging the malignancy degree of ccRCC is given by the Fuhrman grading system, which defines four pathological grades based on nuclear size, shape, staining, and nucleoli (5). Generally, tumors with low invasiveness are classified as grades I–II, and those with high invasiveness are classified as grades III–IV (6).

The preoperative biopsy is the gold standard for evaluating the grade of ccRCC. However, patients undergoing biopsy are at risk of complications, e.g., hematuria (with more than 80% incidence), perirenal hematoma (with about 60%–90% incidence), and infection (7). The procedure of preoperative biopsy is complex and invasive. Besides, biopsy cannot reflect the Fuhrman grade of the whole tumor (8) because of the high spatial and temporal heterogeneity. Thus, preoperative evaluation of ccRCC using a noninvasive procedure for clinical diagnosis is urgently needed. CT examination is the most commonly used non-invasive technique for preoperative diagnosis and follow-up and plays an essential role in diagnosing and treating renal carcinoma, e.g., detection, localization, characterization, and grading of lesions. In some studies, it has been used to evaluate the preoperative ccRCC classification, such as (9–12).

Preoperative noninvasive prediction of ccRCC is conducive to delivering an individualized treatment. Previous studies (12, 13), using radiation characteristics based on multiphase CT, investigated the predictive performance of different machine learning models for discriminating ccRCC. Beyond that (14–17), have shown that convolutional neural networks based on single or multiphase CT images are beneficial for evaluating ccRCC grading. However, the biggest challenge of CT image grading is the existence of noise labels in the image (8). What is a "noise label?" One CT image may contain multiple grades but may have only one label. For example, the grade III–IV grade CT image contains grade I–II tumor areas because the label was obtained from the biopsy of the most severe tumor area of the whole kidney. When encountering noise labels, convolutions uniformly process all tumor regions regardless of their importance, which leads to the inefficiency of classification. CNN makes decisions based on the convolution kernel, which only focuses on the local pixel subset, resulting in the network tending to learn the local mode rather than the global context.

The transformer network, a branch of deep learning, is considered a promising technology for analyzing medical images because it can capture global representations and establish long-distance relationships within the image (18, 19). Therefore, the transformer is suitable for handling CT images with noise labels. We propose a transformer classifier, TransResNet, to predict high-grade ccRCC using CT images. To the best of our knowledge, there have been no investigations about discriminating the low and high nuclear grade ccRCC by combining transformer network and radiological features.

As a result, this research aims to investigate an efficient transformer classifier for predicting the Fuhrman grading of ccRCC based on three-phased CT images.

2 Materials and methods

2.1 Data preparation

Patient cohort: This diagnostic and observational study was approved by the institutional review board (West China Hospital, Sichuan University), and written informed consent was obtained from all patients. Consecutive patients were collected from April 2010 to December 2018 in one hospital. We cleaned the original data according to the following rules: (1) there is no obvious noise in the image of the patient; (2) the patient image has no apparent artifacts. Specifically, all of the pathologically proven ccRCC grades are reconfirmed by experienced radiologists. This work uses the Fuhrman nuclear grading system as the standard grading system. Finally, 759 patients were included in this work.

CT images: All patients underwent a multi-slice CT scan with three phases, including unenhanced, arterial, and portal venous phases, using the following systems: LightSpeed VCT (GE Healthcare), Sensation 64 CT (Siemens), or Sensation 16 CT (Siemens). The PCP, CMP, and NP of the MDCT (multidetector CT) examination were acquired for each patient with the following protocol. By using a high-pressure injector at a rate of 3.5 ml/s, 70–100 ml of contrast agents were injected into the antecubital vein. The CMP is the corticomedullary phase contrast-enhanced scan starting 30 s after the contrast agent injection. The NP is the nephrographic phase contrast-enhanced scan starting 90 s after the contrast materials injection. Spiral scanning and thin-slice reconstruction were used for all three phases. The CT scanning parameters of the phases were as follows: the tube voltage was 120 KV; the reconstruction thickness was 1 mm to 5 mm, and the matrix was 512 × 512. All CT scans of patients are converted to color-scale images and reviewed by experienced radiologists in abdominal imaging. The ccRCC images collected from the CMP phase are used in experiments.

CT Processing: Medical CT slices are diverse and complex. If CT slices are selected to be the input of the classification model, the results are imperfect and require further optimization. To make a more accurate diagnosis, we preprocess the original CT image by detecting the organ or lesion area from medical scanning. We utilize a tumor detection network to quickly and efficiently obtain the rectangular region of interest of the tumor in each phase image. As shown in Figure 1, to reduce the complexity of direct tumor segmentation, the detection frame is divided into two stages: (1) renal organ detection: this detection module is composed of VGG16 (20) without the classification layers (pre-trained on ImageNet (21)), aiming to find the rectangular region of kidneys to mitigate the influence of background of CT scans and reduce the search space of tumors; (2) tumor detection: aims to regress the rectangular region of the tumor accurately and predict the possibility of the tumor.

Because medical data are scarce and difficult to label, large deep learning models rely on data augmentation to improve performance. To study the impact of data enhancement, we will consider several common enhancements here. There are two types of data augmentation in the vision computer domain. One is appearance transformation, such as sharpness, brightness, contrast, saturation, gray processing, Gaussian blur, and elasticity (22), another involves spatial geometric transformation, such as horizontal flipping, rotation, cropping, and resizing (23). Every enhancement strategy can transform data stochastically with internal parameters (e.g., rotation degree, noise level). Our model adopted strategies including random clipping, Affine, Gaussian blur, and Gaussian noise. Figure 2 visualizes the augmentations that we study in this work.

2.2 ccRCC classification network

To make the features extracted from the ccRCC dataset correspond to the label as much as possible, we introduced a transformer module into convolutional networks to improve performance. We illustrate the overall diagram in Figure 3A. Firstly, the input image is processed with several convolution blocks, and then the processed feature map is provided to the transformer block. At the end of the network, the class tokens are applied for the prediction of ccRCC. Our insights into taking advantage of convolutions and transformers are: (1) in the early network, using convolution to learn densely distributed and low-level features requires less computational cost than the transformer; and (2) in the later network, applying the transformer to learn higher-level semantic concepts and long-range dependencies in the image.

The network framework, termed TransResNet, is composed of convolution blocks, transformer blocks, and classifier (one FC layer). The convolution block adopts a pyramid structure, in which the resolution of feature mapping decreases with the increase in network depth while the number of channels increases. It consists of four phases: the first phase is a 3 × 3 convolution with stride 1 and padding 1, which is used to extract initial local features. In the last three stages, we refer to the first three layers of ResNet, in which the output channels of each

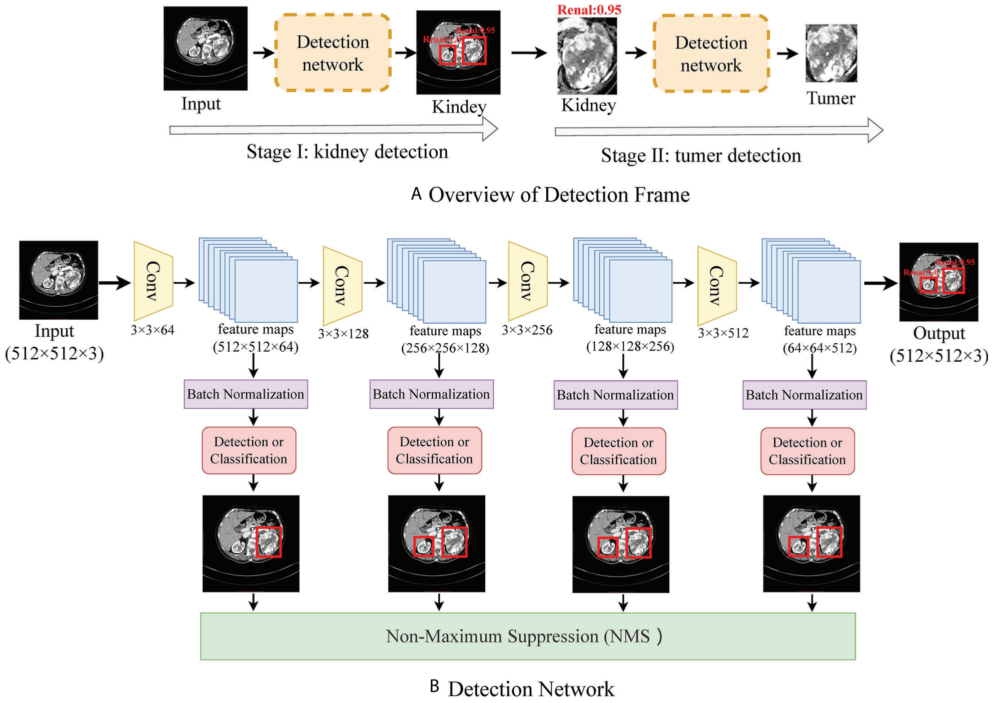
Renal:0.95
Input
Detection network
Kindey
Kidney
Detection network
Tumer
Stage I: kidney detection
Stage II: tumer detection
A Overview of Detection Frame
Input (512×512×3)
Conv
3×3×64
feature maps (512×512×64)
Conv
3×3×128
feature maps (256×256×128)
Conv
3×3×256
feature maps (128×128×256)
Conv
3×3×512
feature maps (64×64×512)
Output (512×512×3)
Batch Normalization
Detection or Classification
Non-Maximum Suppression (NMS)
B Detection Network

FIGURE 1
(A) shows the flow frame of data processing, in which stage one is the kidney detection network and stage two is a tumor detection network. **(B)** shows the details of the detection network.

layer are 16, 32, and 64, respectively. The Transformer block contains 12 repeated transformers. As shown in Figure 3B, each transformer consists of a multi-head self-attention (MSA) module and a multi-layer perceptron (MLP) block.

In the Transformer, we consider the input feature map $X \in \mathcal{R}^{c \times h \times w}$ with c channels and the feature shape of $h \times w$. Self-attention estimates the relationship between one part of a feature map and other parts (e.g., which tumor masses are likely to come together into a complete tumor mass with maximum grade). Therefore, the feature map is divided into a sequence $X=\{x_1,x_2,\ldots,x_n\}$ $(x \in R^{c\times\frac{h}{n}\times\frac{w}{n}})$ with n patches. The goal of self-attention is to capture the interaction among all n patches by encoding each patch in terms of global contextual information. The output of the MSA layer is computed with the feature map X using the following equation:

$$Z = MSA\left(LN\left(softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V + X_{class} + X_{position}\right)\right) + X \qquad (1)$$

where Queries W^Q , Keys W^K, and Value W^v are learnable weights to automatically learn the importance of each patch. The

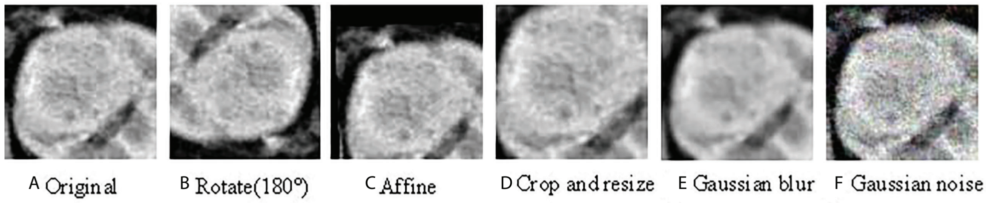

FIGURE 2
Illustrations of the results using various data augmentation **(A)** shows the original CT image; **(B)** is the CT image after 180° rotation; **(C)** is the CT image after Affine transformation; **(D)** shows the CT image after crop and resize transformation; **(E)** is the CT image with Gaussian blur; **(F)** is the CT image with Gaussian noise.

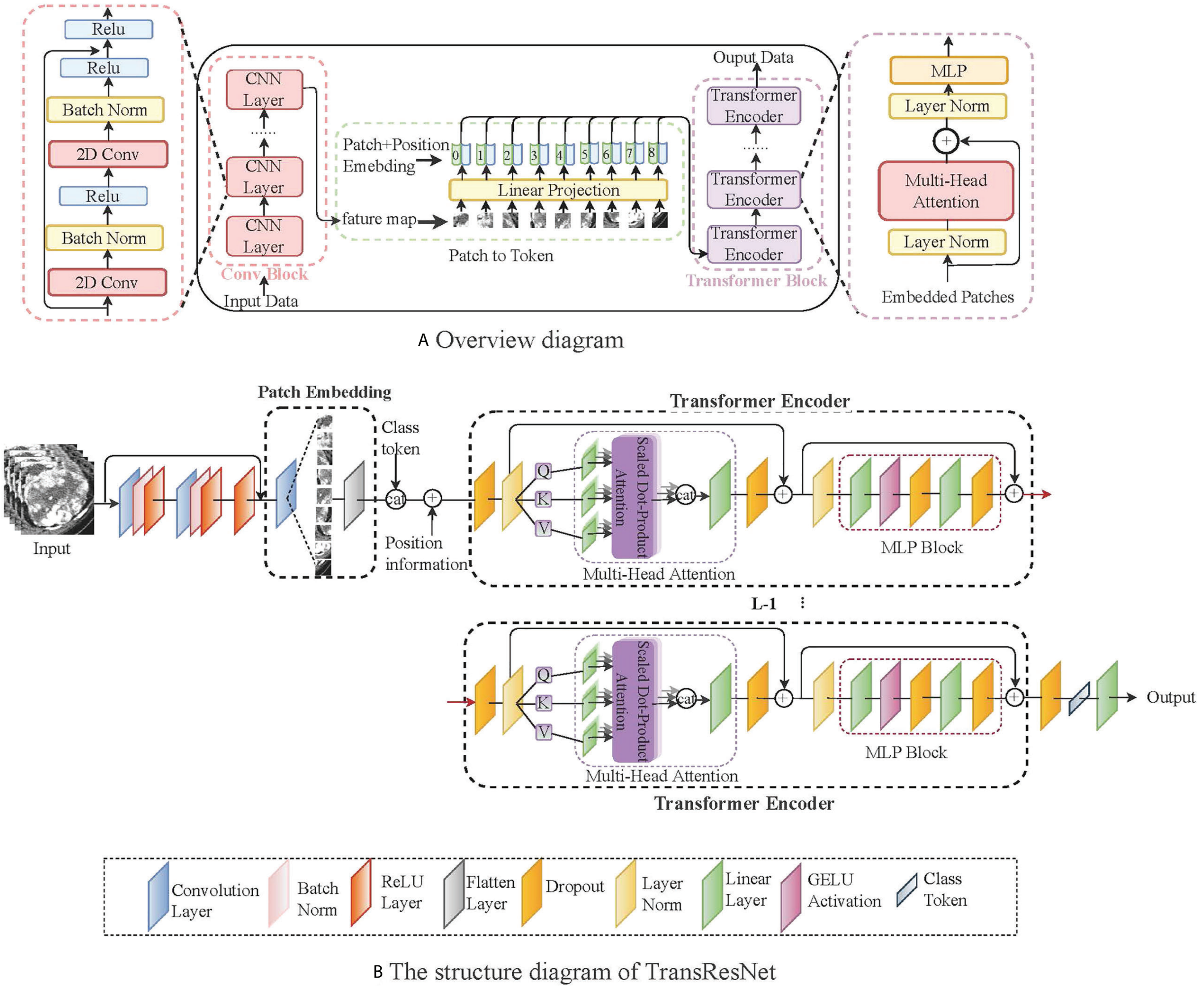

FIGURE 3
A simple network framework for TransResNet. **(A)** represents the overall architecture, mainly including the CNN structure and the transformed structure. **(B)** shows the details of TransResNet. The network framework contains 12 transformer blocks (i.e., L = 12).

input sequence X is projected onto these weight matrices to get $Q{=}XW^Q$, $K{=}XW^K$ and $V{=}XW^V$. The d_k is the dimension of key vector k, providing an appropriate normalization to make the gradient more stable. LN represents the linear normalization, X_{class} and $X_{position}$ are randomized parameters representing classification and position information, respectively. To capture the structure of object image, position and classification information are merged into the self-attention feature map for training.

After the self-attended feature map Z passes through the linear layer, we put it into a MLP including two linear transformations and a GELU activation to perform the same operation on the vector of each position. The output can be obtained:

$$Y = MLP(LN(Z)) + Z \tag{2}$$

Referring to ViT (24), we take the classification information X_{class} instead of extracted image features as the input of the classifier FC to classify directly. We train our network using the cross-entropy (CE) loss between prediction and ground truth, which can be written as:

$$\mathcal{L}_{CE} = -\sum \; (y \log \, \hat{y} + (1-y) \log \, (1-\hat{y})) \tag{3}$$

where the y is the real label and $\hat{y}$ is the predicted probability. The loss function represents the difference between the real label and the predicted probability.

We constructed TransResNet with three residual blocks and 12 transformer blocks. The major hyper-parameters are as follows: the optimizer is Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5e–4. The batch size is 100 per worker. For the epochs, the learning rate is scaled linearly from 0.01 to 0.00001, and then it is divided by 10 at epochs 50, 100, and 150. The proposed model is implemented using Pytorch 1.0.1. We ran the experiments on an Ubuntu 16.04.3 server with four NVIDIA GeForce GTX 1080 cards. Specific details of the code can be seen at: https://github.com/yangmeiyi/ccRCC_project.git.

2.3 Model ensembles

To a certain degree, ensemble learning can improve prediction accuracy. To achieve a strong classifier, we train multiple different classifiers. We selected the following five networks as sub-networks of the integration model:

TransResNet: In medical, ResNet (25) is the preferred model because it is simple and efficient. The most common model, ResNet50, is ineffective on our dataset due to serious over-fitting. Finally, we decided to pair Transformer and resnet34.

TransDenseNet: DenseNet (26) found that some layers are randomly lost at each step in the training process, which can significantly improve the generalization performance of ResNet. Similarly, we chose to combine Transformer and DenseNet-121 with the smallest model parameters as our sub-network.

TransInception: Compared with the structure of ResNet, Inception (27) not only increases the width of the network but also embeds features of different scales. Similarly, Inception-V3 was selected by us.

TransSENet: Different from the previously mentioned networks, improving the performance through spatial latitude, SENet (28) establishes the interdependence between feature channels.

TransRegNet: RegNet (29) proposes a new design paradigm that estimates the overall network design space (depth and width) to obtain the best design. Similarly, we chose RegNetY-200MF with the minor model parameters.

We combine them through the average method. These classifiers and TransResNet are constructed in the same way. The difference lies in the construction of convolution blocks, such as TransDenseNet uses a dense convolutional network and TransInception adopts an inception network.

Suppose we have got N trained different classifiers. Facing a test sample x ($x \in$ test dataset), the prediction results of the integrated model are a N-dimensional vector $\{S_1, S_2, \ldots, S_n\}$. The final score using averaging is formulated as follows:

$$Final\_score = \frac{\sum_{i=1}^{N} S_i}{N} \tag{4}$$

3 Results

3.1 Clinical characteristics

This study included 759 patients, which comprised 477 low-grade [grade I (n = 25, 5.24%) and II (n = 452, 94.76%)] ccRCC patients and 282 high-grade [grade III (n = 232, 82.27%) and IV (n = 50, 17.73%)] ccRCC patients. Male and female patients are equally represented, accounting for 64% and 46%, respectively. The patient characteristics of the training and testing cohorts are shown in Table 1.

3.2 Performance of proposed algorithm

To validate this model, we choose Accuracy (ACC), Area Under Curve (AUC), Sensitivity (SE), and Specificity (SP) as evaluation criteria.

Performance with data enhancement: To comprehend the role of enhancement strategy in detail, the individual and different combinations of them are discussed. With inconsistent images of the ccRCC, cropping as the basic data processing is applied. Figure 4 shows the evaluation results under single and combined transformations. Even though Affine is a very effective enhancement for the model, any single transformation is insufficient for representation learning. One augmentation composition stands out: random Affine and random Gaussian noise. The best combination is more than two enhancements, such as random Rotate, random Affine, and random Crop, making the model obtain 87% ACC and 91% AUC.

Performance on different architectures. This section shows the effect of the transformer on ResNet (25), DenseNet (26), Inception (27), SENet (28), and RegNet (29). Each model has its own unique advantages. We present the ACC, AUC, SE, and SP of TransResNet, TransDenseNet, TransInception, TransRegNet, and TransSENet on the ccRCC dataset in Table 2. Each architecture was performed under the enhancement method of

TABLE 1 The demographic and clinical statistics of patients with ccRCC.

Attribute	Training cohort	Testing cohort
Age (years)	56 ± 12 (589)	54 ± 11 (170)
Male	374	112
Female	215	58
Grade I	21	4
Grade II	371	81
Grade III	165	67
Grade IV	32	18

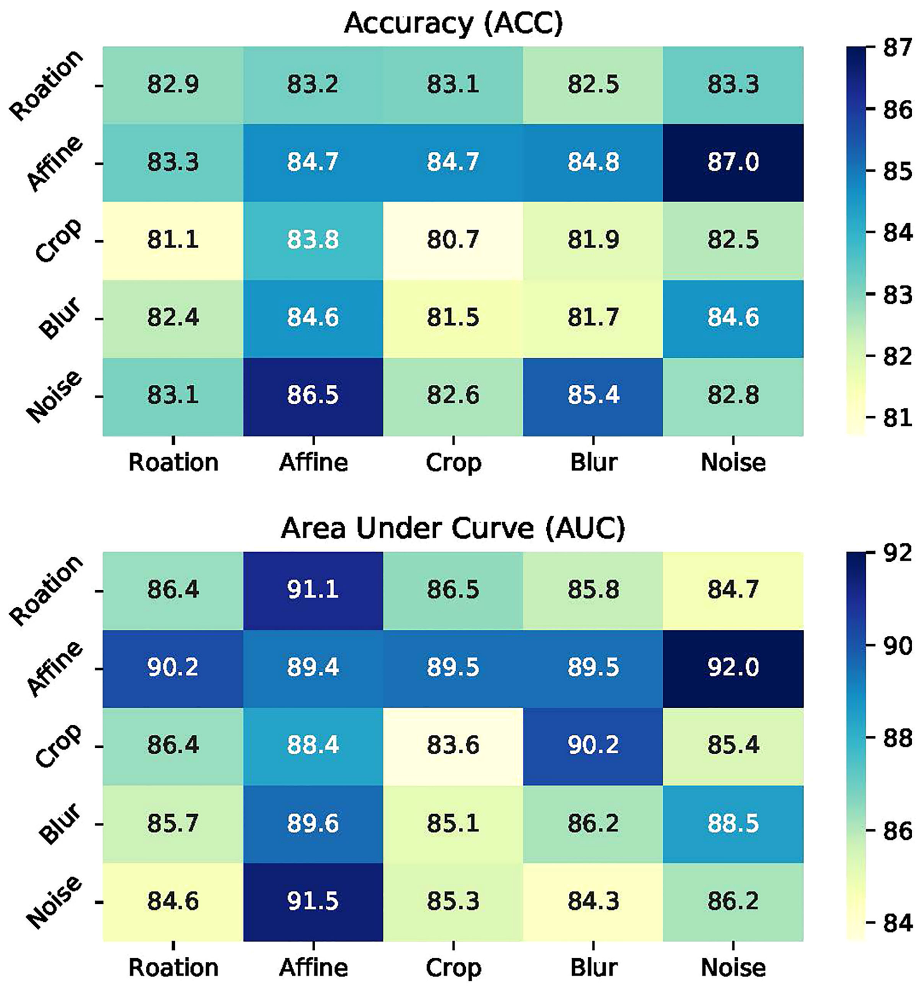

FIGURE 4
Illustrations of data augmentation operators.

Crop, random Horizontal Flip, random Rotation, and random Affine.

Table 2 shows the experimental results of different network architectures. The Receiver Operating Characteristic (ROC) curve is a metric that can provide a pure index of accuracy, which has been accepted by every researcher and applied in medical studies. We show the ROC curve in Figure 5.

Ensemble results: Ensemble learning is a popular way to improve robustness and accuracy by training a group of heterogeneous models. These heterogeneous models are combined by different strategies such as voting, averaging, stacking, and blending. This study analyzes various ensemble strategies, and the results are shown in Table 3. The integrated model consists of a single model: TransResNet, TransDenseNet, TransInception, TransRegNet, and TransSENet.

3.3 Comparison with SOTA methods

Does our proposed solution lead to better performance than pure transformers? To answer this question, we compared pure transformer networks ViT (24) and CaiT (30) with the smallest parameters and a hybrid network Conformer (31) trained in parallel by convolution and transformer in Table 4. In the training process, ViT-Small and CaiT-Small are suitable for our dataset, but the results are still unsatisfactory. In particular, the hybrid structure Conformer with huge model parameters is difficult to train on small dataset ccRCC, resulting in the worst performance, although its classification results on ImageNet are better than ViT and CaiT.

Does the transformer improve performance compared to CNN? For CNN, we have done a series of experiments with ccRCC for comparison. The experiments included Data

TABLE 2 The impact of transformer on different network structures. .

Model	ACC	AUC	SE	SP
ResNet-34	82.3 ± 2.5	85.7 ± 2.3	89.4 ± 1.7	83.2 ± 1.2
TransResNet	87.1 ± 2.3	90.3 ± 2.5	91.3 ± 1.4	85.3 ± 1.5
DenseNet-121	81.5 ± 2.3	85.5 ± 2.4	80.0 ± 0.5	84.3 ± 1.2
TransDenseNet	83.9 ± 2.1	90.5 ± 2.2	80.6 ± 0.6	86.8 ± 1.0
Inception-V3	80.0 ± 2.0	83.7 ± 2.0	76.5 ± 1.2	78.8 ± 1.3
TransInception	84.3 ± 2.0	89.4 ± 2.0	83.4 ± 1.3	85.8 ± 1.6
SENet	81.8 ± 2.5	84.2 ± 2.8	76.8 ± 1.5	85.1 ± 1.4
TransSENet	85.1 ± 2.3	89.2 ± 2.5	89.1 ± 1.3	82.1 ± 1.1
RegNet	81.9 ± 2.1	84.3 ± 2.5	82.5 ± 1.5	80.0 ± 1.0
TransRegNet	82.3 ± 2.0	87.7 ± 2.5	84.7 ± 1.0	80.0 ± 0.5

We report the average accuracy and standard deviation of five time runs.
ACC, AUC, SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively.

Enhancement, CNN architecture, Transfer Learning, and the Regularization Method. The experimental results are shown in Figure 6. In the experiment of data enhancement, the performance of the CNN model has definite improvements after random Affine. However, other data enhancement methods are not effective for the dataset. We found that the CNN architectures have little effect on the Fuhrman Grade of ccRCC, although the advantages of each structure are different. From various experimental results, the model based on CNN is difficult to break through the bottleneck of AUC (e.g., 85 AUC). Our method, TransResNet, easily surpasses CNN algorithms using various tricks in ACC and AUC.

3.4 Comparison with transfer learning

Transfer learning is widely used in medical image processing, such as type 2 diabetes (32), 3D genome folding (33), and papillary thyroid carcinoma (34), etc. Many factors make it impossible to establish large-scale datasets such as ImageNet in the medical field, so the limited data restricts the performance of deep neural networks. Some studies reckon that the pre-trained model obtained on ImageNet can be fine-tuned on the medical dataset to acquire high performance. The experiment by Shin et al. proved that although there are differences between natural images and CT images, CNN fully trained on large-scale well-annotated datasets may still be

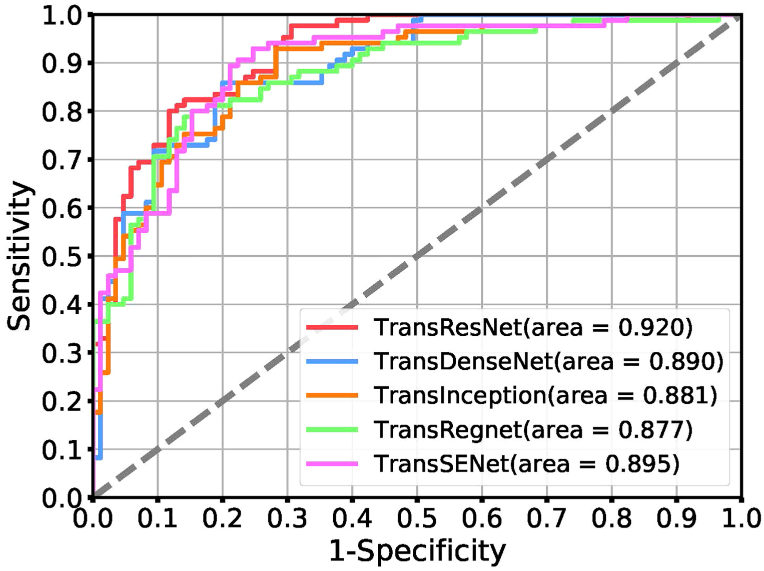

FIGURE 5
Receiver operating characteristic (ROC) curves for the task of tumor classification using a positive ratio feature.

TABLE 3 Ensemble results under different strategies.

Model	ACC	AUC	SE	SP
TransResNet	85.8	90.3	91.3	85.3
TransDenseNet	83.9	90.5	80.6	86.8
TransInception	84.3	89.4	83.4	85.8
TransRegNet	82.3	87.7	84.7	80.0
TransSENet	85.1	89.2	89.1	82.1
Ensemble(voting)	85.8	90.5	91.4	89.8
Ensemble(averaging)	**86.5**	**91.2**	**92.1**	**89.5**
Ensemble(stacking)	86.1	90.8	89.1	91.1
Ensemble(blending)	86.1	90.0	90.0	85.5

ACC, AUC, SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively. The averaging strategy of ensemble shows the best results (the seventh column, highlighted in bold).

transmitted to make the medical image recognition task more effective (35). Therefore, we compared the performance of our method and transfer learning on ccRCC. We transferred the ResNet, DenseNet, and ViT trained with ImageNet to ccRCC for fine-tuning, respectively. In transfer learning, to make the network training more thorough, we follow the method of (35): all CNN layers are fine-tuned with a learning rate 10 times lower than the default learning rate, except for the last layer. The last fully connected layer is randomly initialized and newly trained to adapt to the new object categories in the ccRCC application. Its learning rate remains at the original 0.01. The results are shown in Table 5.

In the experiment, we found that although transfer learning can speed up convergence early in training, it does not improve final accuracy. As shown in Table 5, the accuracy of our proposed method can quickly catch up with the best of transfer learning and stay higher.

3.5 Analysis of noise labels

Previous articles (15, 17) studied the impact of CNN-based deep learning on ccRCC. These studies prove that the features from CT images extracted by CNN can be effectively used for ccRCC grading and obtaining the SOTA performance. Due to the inconsistency between extracted features and labels, some data are always misclassified by the CNN-based algorithm. For example, 21 CT images in our test set could not be classified correctly, including three low-grade images and 18 high-grade images. This phenomenon is consistent with the fact that high-grade images contain more noise information. In other words, the features extracted from tumor images scanned by high-grade ccRCC CT scans are not always related to Fuhrman nucleus grades III and IV, and some of the features are related to Foreman nucleus grades I and II. Figure 7 shows the case of misclassification data, in which Figure 7A is a positive sample, but it is classified as a negative example. Figure 7B is a negative example sample of classification error. Figure 7C is a class activation map of a positive example sample of classification error. Figure 7D is a class activation map of a positive example sample of classification error. We found that: (1) similar CT images with different categories are difficult to distinguish (e.g., Figures 7A, B), and (2) some CT images are not activated correctly (e.g., Figures 7C, D).

The Transformer offers many advantages over traditional CNN algorithms for ccRCC grading. For example, Transformer can capture the representation of global images and establish long-distance relationships in CT images, reflecting the microscopic heterogeneity changes of tumors more comprehensively and providing a more accurate diagnosis than all CNN algorithms. In addition, the complementary strengths between CNN and Transformer can be adapted to the requirements of specific clinical environments. After the optimized classifier TransRsenet, 14 CT images in our test set could not be correctly classified, including two low-grade images and 12 high-grade CT images. Figure 8 shows the class activation map of these error data to show the shift of attention of the model during training. Compared with the convolutional network, the model with the proposed method makes objects clearer and more accurate than the original ones.

TABLE 4 Comparison with the state-of-the-art transformer.

Model	ACC	AUC	SE	SP
TransResNet	**85.8**	**90.3**	**91.3**	**85.7**
ViT-Small	78.6	83.5	88.5	81.3
CaiT-Small	79.4	83.2	89.1	82.0
Conformer	76.4	79.3	83.5	65.7

ACC, AUC, SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively. Our model TransResNet obtains the best results (the first column, highlighted in bold).

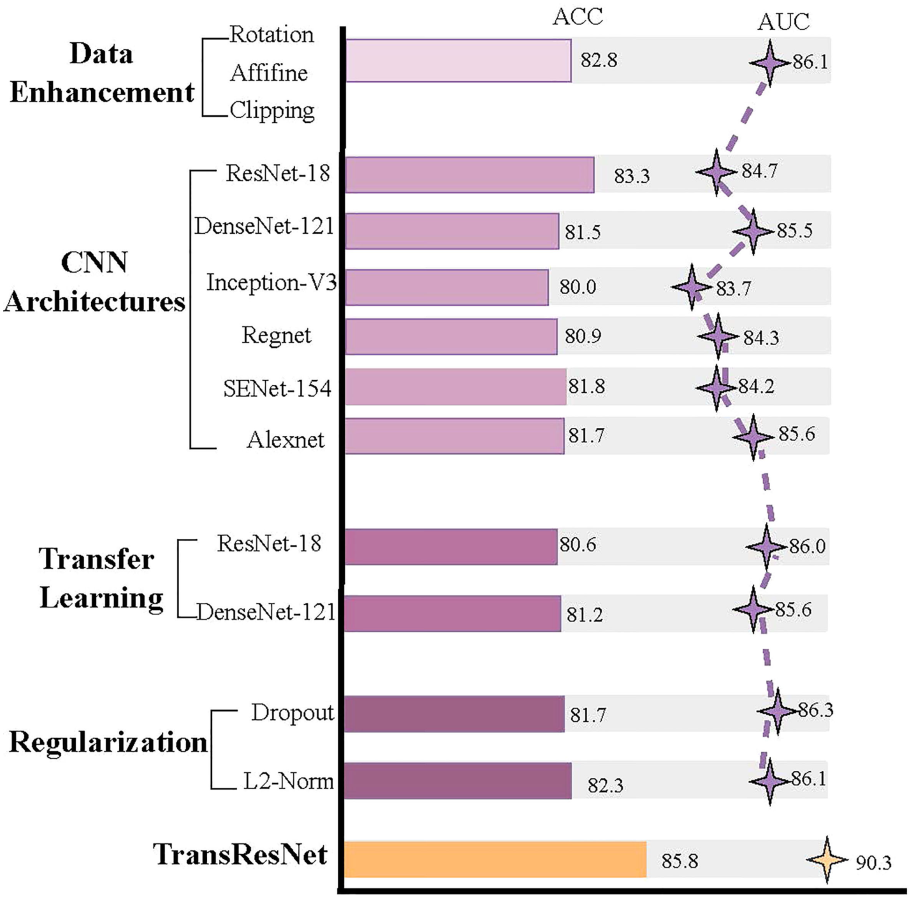

FIGURE 6
A series of experimental results about CNN.

In addition to the comparison of convolution algorithms, we also compared the network architecture of SOTA, such as the transformer-based ViT, CaiT, and Conformer, obtaining the best grading results under ACC, SE, SP, and AUC. In conclusion, we demonstrated the effectiveness of the Transformer module over ccRCC grading. Furthermore, the combination of CNN and Transformer mitigates the noise label problem in ccrRCC.

3.6 External validation

We evaluate our model based on one external validation dataset, the Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) (36, 37). The TCGA-KIRC focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome

TABLE 5 The comparison results of our method and transfer learning.

Transfer Model	ACC	AUC	SE	SP
ResNet18	80.6	86.0	83.3	78.8
ResNet34	80.2	85.6	86.3	78.2
DenseNet121	81.2	85.6	78.9	89.4
TransResNet	**85.8**	**90.3**	**91.3**	**85.3**
ViT-tiny	74.7	75.6	64.7	84.7

The comparison models are ResNet, DenseNet, and ViT-tiny parameterized by pre-training using the ImageNet dataset. Our model is trained from random initialization.
ACC ,AUC ,SE, and SP are the accuracy, the Area Under Curve, the sensitivity and the specificity of the model on the testset, respectively. Our model TransResNet obtains the best results (the seventh column, highlighted in bold).

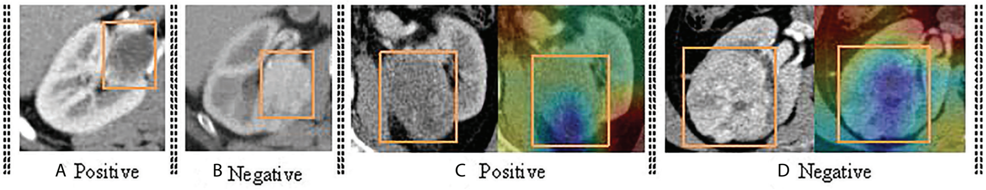

FIGURE 7
The demonstration of error classification. It mainly includes four categories. **(A)** shows the the CT image of positive samples; **(B)** shows the the CT image of negative samples; **(C)** shows the CT image of positive samples and corresponding class activation map; **(D)** shows the CT image of negative samples and corresponding class activation map.

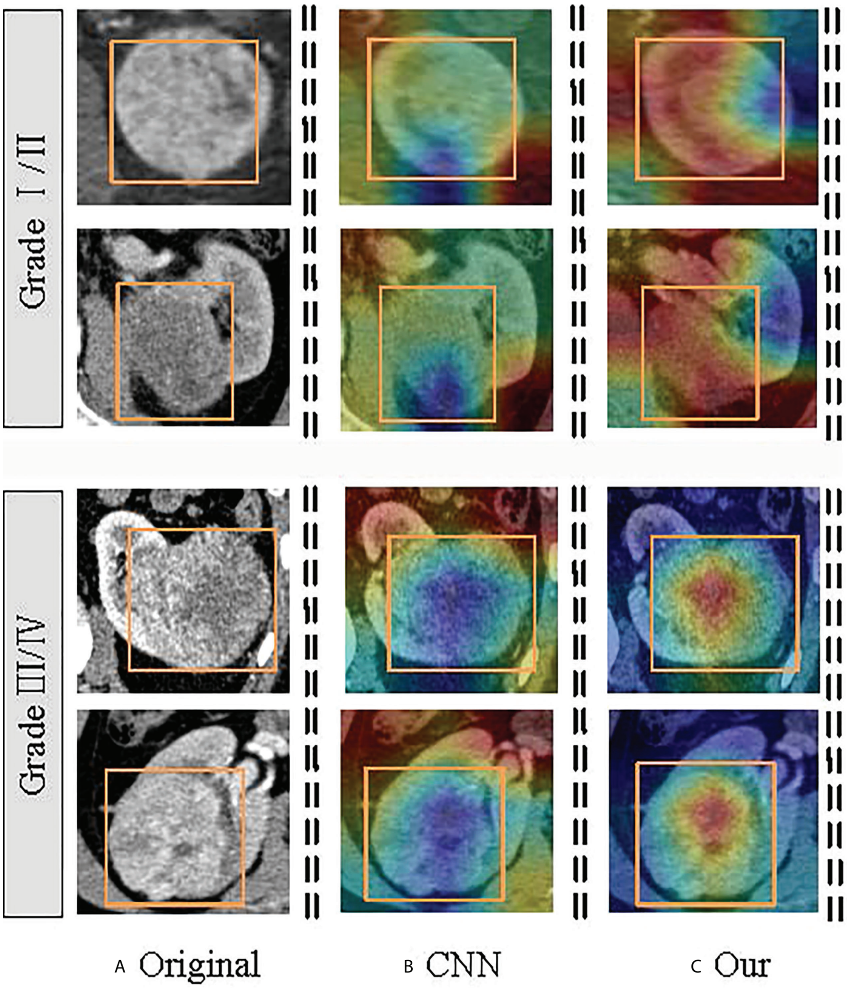

FIGURE 8
Visualization of the class activation map generated by the last transformer layer on images from the ccRCC. The yellow box indicates the lesion area. The color red denotes higher attention values, and the color blue denotes lower. **(A)** is the original image; **(B)** shows the class activation map of the CNN model; and **(C)** is the class activation map of the TransResNet model.

Atlas (TCGA). Clinical, genetic, and pathological data resides in the Genomic Data Commons (GDC) Data Portal, while the radiological data is stored in The Cancer Imaging Archive (TCIA). Here we just use the CT radiological data. We selected 20 patients (e.g., 10 low-grade samples and 10 high-grade samples) from 227 patients as our external validation data. The results of our model TransResNet on the data are: 81.9% ACC, 85.4% AUC, 76.6% SE, and 87.2% SP.

4 Discussion

This retrospective study comprehensively investigates the pros and cons of different deep learning models based on two-phase CT images for discriminating low- and high-grade ccRCC. According to data scarcity and noise labels, a transformer-based deep learning model, TransResNet, is proposed. After a thorough comparison, the proposed discriminative model, TransResNet, achieved satisfactory performance. In addition, we find that this model can effectively alleviate the impact of noise labels.

At present, the analysis of radiomics is a common paradigm to predict the ccRCC. Previous studies analyzed the texture of non-enhanced CT through machine learning to obtain universal features for higher accuracy. For example, Coy et al. have shown that high-grade ccRCC lesions are significantly larger and have more calcifications, necrosis, aggregation system infiltration, and unclear tumor margins than low-grade ccRCC lesions (38). In addition, several studies (39, 40) have proved that high-grade tumors tend to be larger than low-grade. However, our experiment found high-grade lesions are not significantly different from low-grade CT images. This is consistent with the findings of (12) and (13). The performance of image-based quantitative indicators varies on different datasets, and their adequacy needs further verification. In this study, we use deep learning to analyze the CT images, which can automatically discover pixel-level features, supporting a more powerful model.

In ccRCC, some high-grade lesions have low-grade features due to unavoidable sampling bias. It was also interesting to find that convolution-based networks focus on local pixel subsets, which results in a tendency to learn local patterns so that the relationships within images are ignored. As a result, in the complex samples, the convolution kernel cannot effectively make decisions, which leads to unsatisfactory grading accuracy. CNNs uniformly process CT images regardless of their importance. In contrast, the transformer uses a self-attention mechanism instead of a CNN, which can establish long-range dependencies in images. The Transformer has been widely used for medical tasks with high accuracy, e.g., prostatic segmentation (41), delineating the epicardium and endocardium (42), multi-modal medical image classification (18), etc. The Transformer can not only be used for segmentation but also carries an advantage in processing images with noisy labels due to its way of capturing image features. However, transformer networks require a large-scale dataset for training because the transformer lacks inductive bias (24). Thus, scarce medical data leads to severe over-fitting, which may further reduce the reliability of the classifier. Therefore, we take advantage of the CNN and a transformer to extract the features of CT images. The introduced transformer module can help CNN obtain the relationship between the pixel blocks inside the image. We find that this hybrid classifier is superior to a single classifier, with ACC increasing from 83% to 87% and AUC increasing from 85% to 90%.

Correctly grading complex samples is one of the significant indicators for evaluating the strength and stability of a deep learning model, which reflects the ability to extract general features. Our method shows some advantages when faced with noise samples and difficultly graded examples (see *Analysis of noise labels* section). As is known to all, different models have respective preferences for data, and the features learned from a single model are not necessarily reliable. This paper trains a set of heterogeneous models to improve robustness and accuracy by integrating their extracted features, with ACC increasing from 85.8% to 86.5% and AUC increasing from 90% to 91.2%. This study shows that our method can effectively reduce the difficulty of grading caused by noise labels. This method can be widely applied to other medical grading tasks, e.g., the grade of neurogliomas and meningiomas.

Our study has several limitations. First, Fan et al. (12) found that the classifier based on a three-phase CT image is better than that based on a single-phase CT image. Theoretically, non-enhanced CT images can provide additional information for diagnosis. However, this study only used the data from the arterial and portal vein phases due to fewer non-enhanced CT images and inapparent texture information. Additionally, we did not collect additional validation sets because the original DICOM data of the ccRCC are difficult to obtain. However, the superiority of the transformer in the feature extraction can be reflected in the training and testing.

5 Conclusion

This paper studies the application of transformer architecture to ccRCC classification. We first collected a high-quality ccRCC CT scan dataset containing more than 759 patients with pathological proven. Then, a hybrid structure, termed TranResNet, is proposed, which compromises the merits of CNN and Transformer. Unlike other transformer models, TranResNet does not require pre-training on large-scale datasets. Finally, we conducted extensive experiments on ccRCC datasets to verify our method. TransResNet achieves good performance over ConvNets and other related transformer architectures, demonstrating promising results in ccRCC classification. We hope it will help future research on this subject, and it can cooperate with radiologists to classify the ccRCC in an actual clinical situation.

Data availability statement

The original contributions presented in the study are included in the article material, further inquiries can be directed to the corresponding author.

Ethics statement

This diagnostic and observational study was approved by the institutional review board (West China Hospital, Sichuan University). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

MY designed the study and all laboratory studies, performed experiments, analyzed the data, and wrote the manuscript. ML, JD, and XC performed laboratory experiments. XH, YW, QL, and SW collected and analyzed the ccRCC data. LX and FZ provided help in data and statistical analysis. LW and XW participated in data analysis and interpretation. ML and BS oversaw the design of the study and all laboratory studies, data analysis, and interpretation, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2021YGLH213, No. ZYGX2022YGRH016), the Municipal Government of Quzhou (Grant 2021D007, Grant 2021D008, Grant 2021D015, Grant 2021D018), as well as the Zhejiang Provincial Natural Science Foundation of China under Grant No. LGF22G010009.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Capitanio U, Montorsi F. Renal cancer. *Lancet* (2016) 387(10021):894–906. doi: 10.1016/S0140-6736(15)00046-X

2. Incorvaia L, Procopio G, Porta C. Renal Cancer. In: A Russo, M Peeters, L Incorvaia and C Rolfo (eds) *Practical Medical Oncology Textbook*. UNIPA Springer Series., Cham: Springer. (2021) pp.755–74. doi: 10.1007/978-3-030-56051-5_45

3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA: Cancer J Clin* (2021) 71:7–33. doi: 10.3322/caac.21654

4. Fuhrman SA, Lasky LC, Limas C. Prognostic significance of morphologic parameters in renal cell carcinoma. *Am J Surg Pathol* (1982) 6:655–63. doi: 10.1097/00000478-198210000-00007

5. Delahunt B, Eble JN, Egevad L, Samaratunga H. Grading of renal cell carcinoma. *Histopathology* (2019) 74:4–17. doi: 10.1111/his.13735

6. Feng Z, Shen Q, Li Y, Hu Z. Ct texture analysis: a potential tool for predicting the fuhrman grade of clear-cell renal carcinoma. *Cancer Imaging* (2019) 19:1–7. doi: 10.1186/s40644-019-0195-7

7. Ljungberg B, Albiges L, Abu-Ghanem Y, Bensalah K, Dabestani S, Fernández-Pello S, et al. European Association of urology guidelines on renal cell carcinoma: the 2019 update. *Eur Urol* (2019) 75:799–810. doi: 10.1016/j.eururo.2019.02.011

8. Lane BR, Samplaski MK, Herts BR, Zhou M, Novick AC, Campbell SC. Renal mass biopsy–a renaissance? *J Urol* (2008) 179:20–7. doi: 10.1016/j.juro.2007.08.124

9. Cornelis F, Tricaud E, Lasserre A, Petitpierre F, Bernhard J, Le Bras Y, et al. Multiparametric magnetic resonance imaging for the differentiation of low and high grade clear cell renal carcinoma. *Eur Radiol* (2015) 25:24–31. doi: 10.1007/s00330-014-3380-x

10. Chen C, Kang Q, Xu B, Guo H, Wei Q, Wang T, et al. Differentiation of low-and high-grade clear cell renal cell carcinoma: Tumor size versus ct perfusion parameters. *Clin Imaging* (2017) 46:14–9. doi: 10.1016/j.clinimag.2017.06.010

11. Wu G, Zhao Z, Yao Q, Kong W, Xu J, Zhang J, et al. The study of clear cell renal cell carcinoma with mr diffusion kurtosis tensor imaging and its histopathologic correlation. *Acad Radiol* (2018) 25:430–8. doi: 10.1016/j.acra.2017.10.016

12. Lin F, Cui EM, Lei Y. Luo lp. ct-based machine learning model to predict the fuhrman nuclear grade of clear cell renal cell carcinoma. *Abdominal Radiol* (2019) 44:2528–34. doi: 10.1007/s00261-019-01992-7

13. Lai S, Sun L, Wu J, Wei R, Luo S, Ding W, et al. Multiphase contrast-enhanced ct-based machine learning models to predict the fuhrman nuclear grade of clear cell renal cell carcinoma. *Cancer Manage Res* (2021) 13:999–1008. doi: 10.2147/CMAR.S290327

14. Nikpanah M, Xu Z, Jin D, Farhadi F, Saboury B, Ball MW, et al. A deep-learning based artificial intelligence (ai) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic mri. *Clin Imaging* (2021) 77:291–8. doi: 10.1016/j.clinimag.2021.06.016

15. Lin F, Ma C, Xu J, Lei Y, Li Q, Lan Y, 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. doi: 10.1016/j.ejrad.2020.109079

16. Coy H, Hsieh K, Wu W, Nagarajan MB, Young JR, Douek ML, et al. Deep learning and radiomics: the utility of google tensorflow? inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic ct. *Abdominal Radiol* (2019) 44:2009–20. doi: 10.1007/s00261-019-01929-0

17. Han S, Hwang SI, Lee HJ. The classification of renal cancer in 3-phase ct images using a deep learning method. *J Digital Imaging* (2019) 32:638–43. doi: 10.1007/s10278-019-00230-2

18. Dai Y, Gao Y, Liu F. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics* (2021) 11:1384. doi: 10.3390/diagnostics11081384

19. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In: Medical Image

Computing and Computer Assisted Intervention, Strasbourg, France, Cham:. Springer. (2021) 12901:36–46. doi: 10.1007/978-3-030-87193-2_4

20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* (2014) arXiv:1409.1556.

21. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vision* (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y

22. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions, in: *IEEE conference on computer vision and pattern recognition*. Boston, USA: Computer Vision Foundation / IEEE. (2015) pp. 1–9. doi: 10.1109/CVPR.2015.7298594

23. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. *arXiv preprint* (2018) arXiv:1803.07728.

24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale arXiv preprint arXiv:2010.11929. (2020).

25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, in: IEEE conference on computer vision and pattern recognition. Las Vegas, USA: Computer Vision Foundation / IEEE. (2016) pp. 770–8. doi: 10.1109/CVPR.2016.90

26. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks, in: *IEEE conference on computer vision and pattern recognition*. Hawaii, USA: Computer Vision Foundation / IEEE. (2017). pp. 4700–8. doi: 10.1109/CVPR.2017.243

27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision, in: *IEEE conference on computer vision and pattern recognition*. Las Vegas, USA: Computer Vision Foundation / IEEE. (2016) pp. 2818–26. doi: 10.1109/CVPR.2016.308

28. Hu J, Shen L, Sun G. Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, USA: Computer Vision Foundation / IEEE. (2018). pp. 7132–41. doi: 10.1109/CVPR.2018.00745

29. Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P. Designing network design spaces, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: Computer Vision Foundation / IEEE. (2020) pp. 10428–36. doi: 10.1109/CVPR42600.2020.01044

30. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers, in: *2021 IEEE/CVF International Conference on Computer Vision*, Montreal, Canada: IEEE. (2021) 32–42. doi: 10.1109/ICCV48922.2021.00010

31. Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J, et al. Conformer: Local features coupling global representations for visual recognition.,In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada. IEEE. (2021) pp. 357–66. doi: 10.1109/ICCV48922.2021.00042

32. Deng Y, Lu L, Aponte L, Angelidi AM, Novak V, Karniadakis GE, et al. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. *NPJ Digital Med* (2021) 4:1–13. doi: 10.1038/s41746-021-00480-x

33. Schwessinger R, Gosden M, Downes D, Brown RC, Oudelaar AM, Telenius J, et al. Deepc: predicting 3d genome folding using megabase-scale transfer learning. *Nat Methods* (2020) 17:1118–24. doi: 10.1038/s41592-020-0960-3

34. Yu J, Deng Y, Liu T, Zhou J, Jia X, Xiao T, et al. Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nat Commun* (2020) 11:1–10. doi: 10.1530/ey.18.3.15

35. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* (2016) 35:1285–98. doi: 10.1109/TMI.2016.2528162

36. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *J Digital Imaging* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7

37. Akin O, Elnajjar P, Heller M, Jarosz R, Erickson B, Kirk S, et al. Radiology data from the cancer genome atlas kidney renal clear cell carcinoma [TCGA-KIRC] collection. *Cancer Imaging Arch* (2016). Website: https://wiki.cancerimagingarchive.net/display/Public/TCGA-KIRC

38. Coy H, Young JR, Douek ML, Pantuck A, Brown MS, Sayre J, et al. Association of qualitative and quantitative imaging features on multiphasic multidetector ct with tumor grade in clear cell renal cell carcinoma. *Abdominal Radiol* (2019) 44:180–9. doi: 10.1007/s00261-018-1688-8

39. Ding J, Xing Z, Jiang Z, Chen J, Pan L, Qiu J, et al. Ct-based radiomic model predicts high grade of clear cell renal cell carcinoma. *Eur J Radiol* (2018) 103:51–6. doi: 10.1016/j.ejrad.2018.04.013

40. Ishigami K, Leite LV, Pakalniskis MG, Lee DK, Holanda DG, Kuehn DM. Tumor grade of clear cell renal cell carcinoma assessed by contrast-enhanced computed tomography. *Springerplus* (2014) 3:1–7. doi: 10.1186/2193-1801-3-694

41. Han C, Rundo L, Murao K, Noguchi T, Shimahara Y, Milacski ZÁ, et al. Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. *BMC Bioinf* (2021) 22:1–20. doi: 10.1186/s12859-020-03936-1

42. Wu Y, Hatipoglu S, Alonso-Álvarez D, Gatehouse P, Li B, Gao Y, et al. Fast and automated segmentation for the three-directional multi-slice cine myocardial velocity mapping. *Diagnostics* (2021) 11:346. doi: 10.3390/diagnostics11020346