Classification of Kidney Tumor Grading on Preoperative Computed Tomography Scans

Maryamalsadat Mahootiha[1,2(✉)], Hemin Ali Qadir[1], Jacob Bergsland[1], and Ilangko Balasingham[1,3]

[1] The Intervention Centre, Oslo University Hospital, 0372 Oslo, Norway
marymaho@uio.no
[2] Faculty of Medicine, University of Oslo, 0372 Oslo, Norway
[3] Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway
https://www.ous-research.no/interventionalcentre/

Abstract. Deep learning (DL) has proven itself as a powerful tool to capture patterns that human eyes may not be able to perceive when looking at high-dimensional data such as radiological data (volumetric data). For example, the classification or grading of kidney tumors in computed tomography (CT) volumes based on distinguishable patterns is a challenging task. Kidney tumor classification or grading is clinically useful information for patient management and better informing treatment decisions. In this paper, we propose a novel DL-based framework to automate the classification of kidney tumors based on the International Society of Urological Pathology (ISUP) renal tumor grading system in CT volumes. The framework comprises several pre-processing techniques and a three-dimensional (3D) DL-based classifier model. The classifier model is forced to pay particular attention to the tumor regions in the CT volumes so that it can better interpret the surface patterns of the tumor regions to attain performance improvement. The proposed framework achieves the following results on a public dataset of CT volumes of kidney cancer: sensitivity 85%, precision 84%. Code used in this publication is freely available at: https://github.com/Balasingham-AI-Group/Classification-Kidney-Tumor-ISUP-Grade.

Keywords: Kidney cancer · Renal cancer · Deep neural networks · Tumor grading · Classification · CT scan

1 Introduction

Kidney cancer (or renal cancer) is among the most commonly diagnosed visceral malignancies, with a significant annual increase in the incidence- and mortality-rate accounting for 431,288 new cases and 179,368 new deaths in both genders

A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 75–89, 2023.
https://doi.org/10.1007/978-3-031-34586-9_6

in 2020 [1]. Surgical removal is still the most common treatment option for localized kidney tumors. Recently, several other targeted therapies for the treatment of kidney cancer have been introduced to improve patient outcomes and avoid surgical intervention [2,3]. Accurate grading and classification of renal cell neoplasia are essential to provide the optimal treatment option and play a major role in the estimation of patient prognosis. There are several grading systems, with Fuhrman grading being the most widely used one. Recently, there have been doubts about the applicability and prognostic value of Fuhrman grading [4]. In 2012, the International Society of Urological Pathology (ISUP) held a conference to address these issues and proposed a novel grading system known as ISUP grading classification, categorizing renal cell carcinoma (RCC) into four grades namely grades 1, 2, 3, and 4 [5]. It has been shown that a tumor's specific information can be observed pre-operatively from the tumor's appearance on cross-sectional imaging such as computed tomography (CT) scan or magnetic resonance imaging (MRI) [6]. Manual interpretation and quantitative evaluation of radiological data is a laborious and noisy process. In addition, there can be hidden information that the human brain can not perceive from this type of data. For example, microscopic morphological changes associated with histological patterns are crucial in establishing the ISUP grading system.

Over the last decade, several computational methods have been proposed to automate renal cancer classification and staging [7–10]. Deep learning (DL) has been the dominant method because of its advances in finding complex hidden patterns from training data and transforming the input images into abstract features. In most of the studies [7–10], renal whole-slide histology images have been the major source of information about microscopic morphological patterns which are associated to different RCC subtypes such as clear cell RCC, papillary RCC, chromophobe RCC, renal oncocytoma, etc. In contrast, there are several attempts to utilize radiological data for the development of automatic kidney cancer classification [11–15] and staging [16,17]. Many DL-based models were proposed for binary classification differentiating benign and malignant renal tumors from either CT scans [13,15] or MRI [12,14]. S. Han [11] modified GoogleNet [16] for discriminating three major subtypes of RCCs using CT image analysis. N. Hadjiyski et al. [16] adapted the 3D variant of the inception model to predict cancer staging, while M. A. Hussian et al. [17] proposed an automatic low stage (I-II) and high stage (III-IV) RCC classification both from CT scans.

In this paper, we propose a novel DL-based framework to computationally classify kidney tumors into ISUP grades from pre-operative CT scans available for each patient. To the best of our knowledge, our work is the first study to investigate DL in 3D images for renal tumor ISUP grading indicating the histopathological patterns and associated with risk score [5,18], which is the basis of the survival analysis. We do not intend to detect and segment the kidneys and the tumors in CT volumes; instead, we assume that the kidneys and tumors are already localized and segmented. We first extract the kidneys from a 3D CT volume using the provided manual ground truth of kidneys. We concatenate the extracted kidneys and the corresponding ground truth of the tumors into a single tensor on the channel dimension. This concatenation step aims to force the DL-based classifier

model to pay particular attention to the surface patterns of the tumor regions. The concatenated tensor is then fed into a three-dimensional (3D) convolutional neural network (CNN) to classify the kidney image(s) into 4 ISUP grades in every CT volume. We adapt EfficientNet [19] as our classifier model. More specifically, we transform the 2D EfficientNet-B7 to its 3D variant so that it can handle 3D volumetric data. We apply various data augmentations to overcome the class imbalance issue and feed the training model with more samples and several preprocessing methods to standardize the inputs and improve their quality. We show that our proposed framework can provide promising results on unseen CT volumes. This initial result can further be developed into a prognosis model, survival analysis, and treatment management plan.

2 Related Work

Recent studies have focused on the automated grading of clear cell renal cancer carcinoma (ccRCC). Some of them employed histopathology images, while others used CT or MRI for this prediction.

For histopathology images, Tian et al. [20], and Yeh et al. [21] employed Fuhrman's grading system to identify ccRCC as low or high grade. In both studies, the authors, in collaboration with pathologists, examined whole-slide images, identified regions of interest (ROIs), and assigned a grade to each ROI. Then, features of histopathology images were retrieved from ROIs for the model training. Tian et al. tried to find an optimal way between a neural net, random forest, and support vector machine for the model training, while Yeh et al. tried to use a support vector machine to train the classifier model. Both Tian et al. and Yeh et al. could get high sensitivity, specificity, and AUC for their models, and they recommended predicting ISUP grades in future work.

For radiological imaging, Sun et al. [22] developed the support vector machine-based method to determine the ISUP grade of kidney cancer with clear cells. In this research, CT images were divided into two categories: low and high grade. Resampling and a Gaussian filter were utilized for denoising at the preprocessing level. Then, the greatest cross-section picked by radiologists was utilized as the ROI. Sun et al. used a feature extraction mechanism to generate three distinct predictive models, each of which was based on a distinct selection of features. The AUC for the third model was the highest at 0.91. Another research study led by Cui [23] graded ccRCC using CT and MR based on the decision tree. Normalization and pixel resampling were utilized for preprocessing level in this research. The classification was determined by low and high ISUP grades. The RoI was determined based on the tumor-containing slices. Cui et al. next attempted to extract the texture of the slices and create features from them; they employed a decision tree to predict a low or high ISUP grade and attempted to test the model using ACC. For Cui et al. model, an ACC greater than 0.70 was achievable. In a different study, Zhao et al. [24] classified MRI images based on ISUP and Fuhrman grading as low or high grade using CNN. This study was a binary classification: low and high grade. Data augmentation

was utilized prior to the model training. The model with the highest AUC was selected as the ultimate model. The model was developed using ResNet50 and 2D CNN. Zhao et al. combined t1 and t2 sequences for the model and included clinical variables such as age, gender, and tumor size in the network design. For low and high ISUP classification, they could gain 0.92 in sensitivity and 0.78 in specificity. These studies [22–24] agree that CT texture analysis can predict ccRCC pathologic nuclear grade noninvasively.

Multiple factors make our methodology superior to that of previous research studies. First, it is based on CT images, which is a non-invasive method; second, it uses deep learning and does not require feature extraction; third, it uses 3D images and 3D models for predicting, so we do not lose any information by changing it to a 2D based model; fourth, we attempted to have four output classifications rather than a binary classification; and finally, we do not employ clinical data in addition to the CT images, therefore our prediction is solely based on the CT scans and does not require any further information.

3 Methods and Materials

Figure 1 illustrates our proposed DL-based framework developed for kidney tumor grading classification based on the ISUP grading system. Every step will be explained in detail in the following sections.

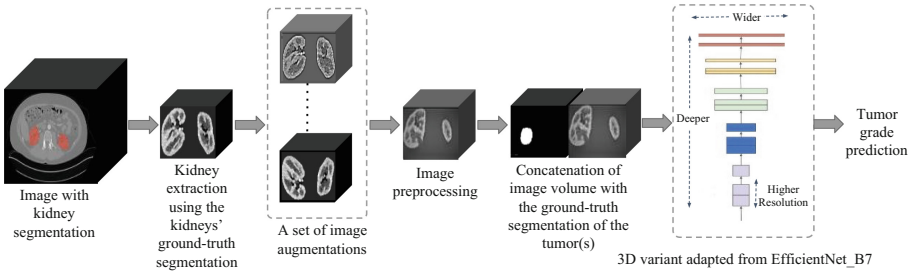


Fig. 1. Overview of the framework proposed for kidney image classification based on the ISUP grading system. We separate the left and right kidneys from the whole image slices during image preparation. We enlarge the number of samples in the training dataset by applying various forms of data augmentation strategies. We enhance the data quality by improving image quality, resizing, and re-orienting the volumes in the image pre-processing phase. To force the model to focus on the tumor surface patterns, we concatenate the image and manual segmentation of the tumors, and finally, we train the classification model with concatenated volumes. The model produces probability values for the four different ISUP grades as the output decision.

3.1 Classifier Architecture

The state-of-the-art convolutional neural network (CNN)architecture for image classification is called Efficient-Net [19]. In a quick but efficient way, Efficient-Net scales up models using the compound coefficient method. The authors of

EfficientNet proposed seven models of various dimensions, which exceeded the state-of-the-art accuracy of most CNNs and had a far higher degree of efficiency. The largest Efficient-Net model, Efficient-Net B7, obtained the best performance on the ImageNet and the CIFAR-100 datasets. The number of parameters in Efficient-Net B7 is higher than the other variants (e.g., B0, B1, B2, B3, B4, B5, and B6). In this study, we adapt the exact structure of Efficient-Net B7 and transform it to a three dimensional (3D) CNN model so that it can handle 3D image data such as CT volumes.

3.2 Dataset

In this paper, we use KiTS21 dataset [25] for training and testing our proposed method. This dataset consists of 300 different patients, each with clinical data and a CT scan with manually annotated kidneys and tumors (ground-truth labels). Patients receiving a partial or complete nephrectomy for suspected kidney cancer between 2010 and 2020 at either the M Health Fairview or Cleveland Clinic medical facility have been included in this dataset. Before surgery, all patients underwent a contrast-enhanced CT scan showing both kidneys. The primary purpose of gathering this dataset was to apply segmentation algorithms.

We attempt to use this dataset since it has a detailed clinical dataset, precise annotation, and adequate subjects. The dataset contains three files as follows: CT scan volumes, annotation volumes, and clinical data. All of the images are in NIFTI format. Each annotation volume contains manual segmentation of the kidneys, tumor(s), and cyst(s). Clinical data is in a JSON file format with 63 fields of clinical parameters for every patient. All essential clinical information, like pathology results, is stored in this file [26]. Originally this data came from the Cancer Imaging Archive, where the imaging and segmentation were stored in DICOM format, and the clinical data was a single CSV file[1].

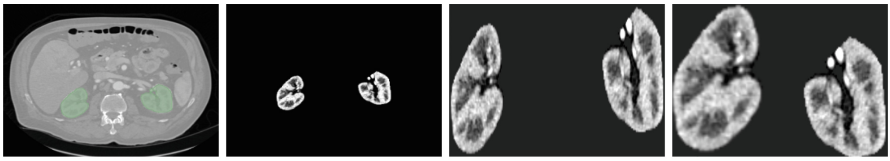
3.3 Data Preparation

Data preparation is a pre-processing mechanism to structure, manipulate, and organize raw data to the data format that the training model can analyze more efficiently. In this study, we apply data preparation on the CT scan volumes and their corresponding annotation volumes.

Image Preparation. The 3D image data from the KiTS21 dataset depict the whole abdomen. The kidneys with tumors cover only a small percentage of the entire image slices. In this study, we aim to train our proposed framework to view only the imaging information related to the kidneys and the tumors. Therefore, we extract the left and right kidneys from the image volumes using the provided ground-truth annotation. Figure 2 shows the steps used to prepare the training samples by removing other organs and extracting the two kidneys from whole image volumes.

[1] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61081171.

The image and segmentation volumes are stored as 3D arrays. The numbers in the image arrays are between 0 and 255, where 0 is the darkest part and 255 is the brightest part. The numbers in the segmentation arrays are 0, 1, 2 and 3 (1 = kidney, 2 = tumor, and 3 = cyst). For image preparation, we have 3 phases. In phase 1, we keep the numbers in image arrays that the corresponding segmentation is one and change the other numbers in image arrays to zero. This step will change all image space except two kidneys to black. Then in phase 2, we keep two kidneys and delete the black background. In phase 3, the black space between two kidneys should also be eliminated, as our first purpose was not to enter the black space into the training model. So we extract the left and right kidneys and merge them again in the width dimension in phase 3.



(a) KiTS21 Image (b) output of phase 1 (c) output of phase 2 (d) output of phase 3

Fig. 2. Image preparation process

Label Preparation. From the clinical dataset file that comes with the KiT21 dataset, we use the tumor_isup_grade field as the label for image classification. This clinical parameter has four values: 1, 2, 3, and 4. ISUP grade of the tumors was indicated in the post-operative surgical pathology report. We notice that in 56 cases, the value of Null is used where ISUP grade does not apply, such as benign tumors or Chromophobes. We remove those 56 patients from our training and testing dataset, leaving us with 244 samples in our final dataset.

3.4 Data Augmentation

The deep learning models frequently need a large amount of training data, which is not always available, in order to make accurate predictions. We apply data augmentation to increase the number of samples in the training dataset. After eliminating those patients without ISPU values, we are left with 244 samples. Patients with the ISUP1 class make 13% of the total, the ISUP2 class 48%, the ISUP3 class 27%, and those with the ISUP4 class 12%. This class imbalance leads to biasing impact on the model training and the final results—the trained model will be more biased toward the dominant class in the training dataset and show poor performance on the minor class. Another challenge that we encounter is that we have to train our classifier model from scratch as we are unable to apply transfer learning or fine-tune a pre-trained 3D EfficientNet-B7 transformed from the original 2D EfficientNet-B7 in this study. Two hundred forty-four samples might not be enough for training a deep neural network for an image classification

model from scratch. A huge quantity of labeled training images is needed for deep learning models to be trained from scratch. We try to partially overcome these two problems with data augmentation.

When strategies like undersampling, oversampling, and data augmentation are used to fix the class balance issue, the model's efficacy increases [27,28]. We don't use oversampling as this method can lead to the model being overfitted to the minority class [28]. Additionally, we avoid using undersampling since we lack sufficient samples in the dataset and don't want to lose any data. As we can see in the literature, performance progress slowed down after 150 images in each class, and after 500 images in each class, there was no noticeable improvement [29]. We found that 500 images per class are enough to attain a reasonable classification accuracy. We increase the number of samples to 2000, 500 in each class. We calculate the number of subjects in each class and realize that class 1 would need to be augmented eleven times, class 2 twice, class 3 five times, and class 4 fourteen times.

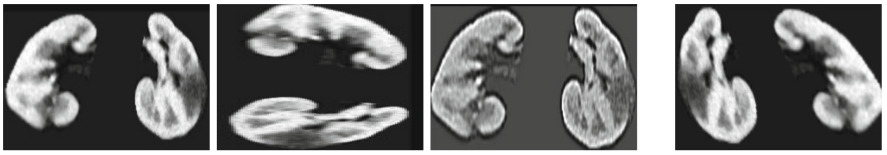
For data augmentation, we do not employ generative adversarial networks and would rather use traditional approaches. The critical point is that if we want to do the augmentation for class4 fourteen times, we must make fourteen different augmented versions of the original data. We use MONAI transformers for data augmentation because the MONAI module is a comprehensive python library for manipulating 3D data such as volumetric images. MONAI library contains all recommended image augmentation techniques to enlarge the number of training samples. Table 1 shows the various transformations we use for data augmentation. We utilize the various combinations of transformers from Table 1 for data augmentation. Figure 3 displays one slice of the original patient's data along with three augmented versions of that slice.

Table 1. MONAI transformers used for data augmentation

Position Augmentation	Noise Augmentation
Affine	GaussianNoise
Rotate90	GaussianSmooth
Flip	GaussianSharpen
	GibbsNoise
	SpaceSpikeNoise

3.5 Data Splitting

Our augmented dataset consists of various copies of the original samples. To prevent unfair performance evaluation of our proposed framework, we split the dataset based on the patient ID into training and testing subsets. In this way, we avoid having the same patient with all its augmented versions in both the training and testing subsets. We use the K-fold cross-validation technique to split out the dataset. We use 3-fold cross-validation. We split our dataset randomly into three



(a) Original image (b) 90° Rotation (c) Gibbs Noise (d) Horizontal Flip and Affine

Fig. 3. Comparison between one axial slice of original image with 3 different augmented versions

different subsets: 162 samples in the training subset; 82 samples in the testing subset. We choose 10% (16 samples) of the training subset as our validation subset in every fold. As we intend to have the same ISUP class distribution in the validation subset, we select four samples from each ISUP grade class.

3.6 Image Pre-processing

Image pre-processing is an essential step before image classification. The purpose of pre-processing is to enhance the image's quality and modify a few of its features so that the training model can better interpret the input [30,31]. We resize all the volumes to $128 \times 128 \times 128$ to have the same size volumes for training the model. We follow the recommended size by the MIT challenge[2] to make the data more manageable. We do not select a bigger size like 256 because a larger image resolution is expensive both in terms of computational power and memory [30]. One millimeter isotropic voxel size is chosen for every volume. This is the standard voxel size recommended by previous studies [32,33]. We re-orient all volumes to the RAS (Right, Anterior and Superior). This is the most common orientation used in medical images [32–34]. We use intensity normalization based on the Z-score in medical imaging [30,35]. We use the image contrast part in ITK snap software[3] for this normalization. We showed the images to the clinicians to identify which contrast range between the kidney and the tumor was more noticeable. So we can figure out the minimum and the maximum contrast number in which the tumor is more distinctive from the kidney. We change intensity values in the image arrays based on this image contrast range.

For kidney image and tumor segmentation, we utilize identical image pre-processing transformers; however, we do not apply intensity normalization for tumor segmentation because the contrast of the segmentation image is not important for training the model.

3.7 Kidney and Tumor Concatenation

In this study, our goal is to classify kidney tumors based on distinguishable surface patterns. To force our 3D EfficientNet-B7 to pay particular attention to

[2] http://6.869.csail.mit.edu/fa17/miniplaces.html.
[3] http://www.itksnap.org/pmwiki/pmwiki.php.

the surface patterns on the tumors, we concatenate the extracted kidneys with their corresponding provided manual segmentation of the tumors. In addition, this image concatenation enriches the input volume with the location and size of the tumors. If we train our 3D EfficientNet-B7 on the kidneys only without providing the location of the tumors, the model may look at other parts of the input volumes and find other patterns and associate them to the classes. This leads to poor performance on unseen data.

3.8 Training Details

We use the Pytorch library for training our model. The experiments are executed in the Linux Ubuntu Operating system on a machine with AMD Ryzen 7 5800X 8-Core Processor, NVIDIA GeForce RTX 3090 GPU and 32 GB RAM. Based on the three folds we previously acquired, we train our model three times but with the same hyperparameters. Each time validation set contains around 10% of the training set. During training, none of the samples from the validation sets are utilized to determine the loss function or back-propagate gradients across the network.

After every training epoch, the model is evaluated on the complete validation set, and the mean AUC[4] is calculated. Model parameters are stored, overwriting the previous model, each time a new best mean validation AUC is obtained. In this regard, compared to all training epochs, the final model that is created during training has the greatest mean validation AUC. We decide on 50 as the number of epochs since we see that the training losses stop decreasing after about 50 epochs. Each model is trained with the help of the Cross-Entropy loss, which is given by:

$$L = -\sum_{i=1}^{n} t_i \times log(p_i), \tag{1}$$

where t_i is the true label and p_i is the softmax probability for the ith class and n is the number of classes.

The ADAM optimizer [36] is used to train the models, and a learning rate of 1×10^{-4} is used since it is empirically proven to produce the best results on clean data [37]. Ten batches are selected to train the model based on the image sizes and computing memory.

4 Results and Discussion

After training the model on the three folds, we evaluated the model's performance. Precision, Recall, and F-score metrics were used to quantitatively evaluate the performance of the proposed framework. The performance metrics are computed from the following formulas:

[4] Area under the curve is a performance measurement for the classification problems. It tells how much the model is capable of distinguishing between classes.

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$\textbf{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

TP is the number of samples that are truly classified, FP is the number of samples that should be in an ISUP class except for the ISUP class-x, but they belong to ISUP class-x; and FN is the number of samples that should be in ISUP class-x, but they are in the other ISUP classes.

Precision, Recall, and F-score was computed for each ISUP class. We calculated the average of four Precision, Recall, and F-scores we gained for each ISUP class. We repeated this process three times for each of the three folds we had, giving us three average Precision, Recall, and F-scores. For our model, we obtained a total Precision of 0.74, Recall of 0.71, and F-score of 0.72 by calculating the mean three average Precision, Recall, and F-scores. Table 2 shows the performance metrics in fold two in which the best performance was obtained.

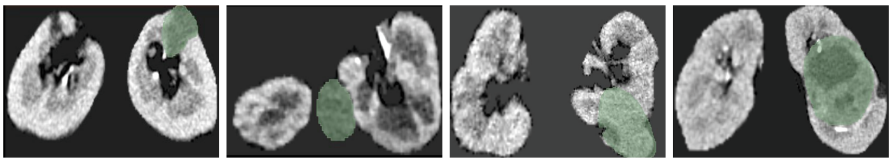
Table 2. Fold 2 performance evaluation of the proposed framework

	Precision	Recall	F-score
ISUP1 class	0.86	0.91	0.88
ISUP2 class	0.79	0.78	0.78
ISUP3 class	0.87	0.77	0.81
ISUP4 class	0.86	0.94	0.89
Average	0.84	0.85	0.84

According to Table 2, the F-scores are high in the following order: ISUP4 class, ISUP1 class, ISUP3 class, and ISUP2 class. If we look back at how many times we augmented the classes, they are high in this order: fourteen times for the ISUP4 class, eleven times for the ISUP1 class, five times for the ISUP3 class, and twice for the ISUP2 class. We can assert that higher accuracy metrics are obtained from a class when there is more augmentation in that class. It arises because the predicted classes for augmented images are frequently the same as those for the original patient image. Most of the time, if the ISUP class of the original image could be accurately recognized, it could also be accurately detected for the augmented version.

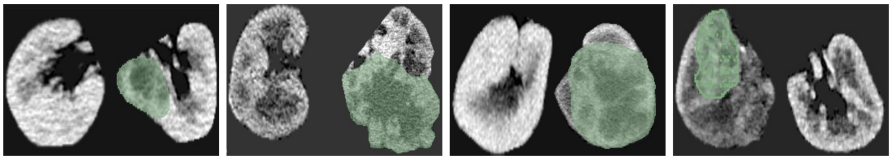
It may be beneficial since it demonstrates how the model can recognize that the augmented image is another version of the original image and forecast the same ISUP class for it. If we look at the accuracy metrics for the ISUP2 class, they are at their lowest, where data augmentation was used the least compared to the other ISUP classes.

Figure 4 illustrates four images from different ISUP classes that are truly classified, and Fig. 5 illustrates four images that are falsely classified. In Fig. 5, we wrote the true ISUP classes as the caption, and the predicted ISUP classes from left to right are ISUP2, ISUP4, ISUP4, and ISUP2. The green parts of the images are the tumor parts. In Fig. 5b, despite the large tumor size, the true ISUP grade was two, and the model identified ISUP 4 in this image. In Fig. 5d, the tumor size was small; the true ISUP class for this image was four, but the model predicted ISUP 2. It demonstrates the model's attempt to concentrate on tumor sizes in its prediction.



(a) ISUP1 class (b) ISUP2 class (c) ISUP3 class (d) ISUP4 class

Fig. 4. Correctly classified images



(a) ISUP1 class (b) ISUP2 class (c) ISUP3 class (d) ISUP4 class

Fig. 5. Misclassified images

Based on a few tumor features, the ISUP grade is determined. When you ask a physician to determine the ISUP class based only on observing CT scan images, they are unable to do so with high certainty [5,18]. We attempted to create a model that could look at patients' CT scans and forecast their ISUP classes. We can conclude that our model was able to extract hidden features relevant to ISPU classes that might not be seen by human eyes.

It is worth motioning that this study has some limitations: 1) to predict ISUP grade, our model needs to get information as input from both the two kidneys and manually segmented tumor(s) indicating the location of the kidney tumors. There is an extract pre-processing stage that extracts the kidneys from the input volume using the manual segmentations of the two kidneys. Our proposed framework might not be able to produce highly accurate classification results from the whole abdominal volumes, and 2) we noticed that sometimes our trained model tries to predict ISUP classes by looking at the tumor size. This impurity leads to ISUP misclassification, so small tumors with grade 4 surface patterns might be classified as grade 1 or 2.

5 Future Work

We can apply transfer learning to improve the performance results [38,39] by getting more three-dimensional images from the cancer imaging archive. Any 3D medical imaging, such as an MRI of the brain or liver, can be used, but for better outcomes, kidney images should be included [40].

Furthermore, we can utilize our image classification layers before fully connected layers as feature extractors since we can use convolutional neural networks for feature extraction [41]. We can link these features to the survival features since the outputs of our image classification are related to the risk score. Thus, we would provide the survival features as the input to the DL-based survival functions, and we can estimate the time of the patient's death by using the patient's medical images.

6 Conclusion

In this study, we proposed a classification framework for kidney tumors based on the International Society of Urological Pathology (ISUP) grading system. We transformed 2D EfficientNet-B7 into a 3D variant that can handle 3D data volumes. To enhance the classification performance, we applied various data augmentation and pre-processing methods. We eliminated other organs in the volumes and kept only the kidneys. The extracted kidneys were concatenated with the provided manual ground-truth annotations of the tumors. This image concatenation is shown to be an important step to force our 3D EfficientNet-B7 to look particularly at the tumors' surface patterns and associate them with the ISUP classes. The data augmentation was applied to first increase the number of samples in the training set and second to partially solve the class imbalance issue. Several image pre-processing methods were applied to enhance the input image quality. The proposed framework demonstrated good classification accuracy of (84%) on the test set. This study shows how crucial it is to properly prepare the dataset through actions like cropping, augmentation, and pre-processing. It is worth mentioning that we tried to show how the results of this work can be generalized to other datasets as well. However, we could not find any similar dataset in which we could get the required information, such as MRI or CT images of the organs, organ segmentation, tumor ground truth, and, most importantly, ISUP grades.

Acknowledgement. This research was funded by the research council Norway and ICT:5G-HEART. We thank our colleagues Davit Aghayan and Egidijus Pelanis from Intervention Center, Oslo University Hospital, who provided insight and expertise in medical images and clinical aspects that greatly assisted the research. We gratitude Piotr Bialecki, Senior Engineering Manager of the PyTorch Team at NVIDIA, for assistance with technical parts of the training model with deep neural networks; and Håvard Kvamme, previous Ph.D. student of the University of Oslo in the faculty of Mathematics, for his comments and showing the actual path of the research.

References

1. Sung, H., et al.: Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **71**(3), 209–249 (2021)
2. Molina, A.M., et al.: A phase 1b clinical trial of the multi-targeted tyrosine kinase inhibitor lenvatinib (e7080) in combination with everolimus for treatment of metastatic renal cell carcinoma (RCC). Cancer Chemother. Pharmacol. **73**(1), 181–189 (2014)
3. Motzer, R.J., et al.: Dovitinib versus sorafenib for third-line targeted treatment of patients with metastatic renal cell carcinoma: an open-label, randomised phase 3 trial. Lancet Oncol. **15**(3), 286–296 (2014)
4. Samaratunga, H., Gianduzzo, T., Delahunt, B.: The ISUP system of staging, grading and classification of renal cell neoplasia. J. Kidney Cancer VHL **1**(3), 26 (2014)
5. Warren, A.Y., Harrison, D.: WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: standards and controversies. World J. Urol. **36**, 1913–1926 (2018)
6. Rees, M., Tekkis, P.P., Welsh, F.K., O'rourke, T., John, T.G.: Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients. Ann. Surg. **247**(1), 125–135 (2008)
7. Zhu, M., Ren, B., Richards, R., Suriawinata, M., Tomita, N., Hassanpour, S.: Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. Sci. Rep. **11**(1), 1–9 (2021)
8. Abdeltawab, H.A., Khalifa, F.A., Ghazal, M.A., Cheng, L., El-Baz, A.S., Gondim, D.D.: A deep learning framework for automated classification of histopathological kidney whole-slide images. J. Pathol. Inform. **13**, 100093 (2022)
9. Abu Haeyeh, Y., Ghazal, M., El-Baz, A., Talaat, I.M.: Development and evaluation of a novel deep-learning-based framework for the classification of renal histopathology images. Bioengineering **9**(9), 423 (2022)
10. Fenstermaker, M., Tomlins, S.A., Singh, K., Wiens, J., Morgan, T.M.: Development and validation of a deep-learning model to assist with renal cell carcinoma histopathologic interpretation. Urology **144**, 152–157 (2020)
11. Han, S., Hwang, S.I., Lee, H.J.: The classification of renal cancer in 3-phase CT images using a deep learning method. J. Digit. Imaging **32**(4), 638–643 (2019)
12. Xi, I.L., et al.: Deep learning to distinguish benign from malignant renal lesions based on routine MR ImagingDeep learning for characterization of renal lesions. Clin. Cancer Res. **26**(8), 1944–1952 (2020)
13. Baghdadi, A., et al.: Automated differentiation of benign renal oncocytoma and chromophobe renal cell carcinoma on computed tomography using deep learning. BJU Int. **125**(4), 553–560 (2020)
14. Nikpanah, M., 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 **77**, 291–298 (2021)
15. Zhou, L., Zhang, Z., Chen, Y.-C., Zhao, Z.-Y., Yin, X.-D., Jiang, H.-B.: A deep learning-based radiomics model for differentiating benign and malignant renal tumors. Transl. Oncol. **12**(2), 292–300 (2019)
16. Hadjiyski, N.: Kidney cancer staging: deep learning neural network based approach. In: 2020 International Conference on e-Health and Bioengineering (EHB), pp. 1–4. IEEE (2020)

17. Hussain, M.A., Hamarneh, G., Garbi, R.: Renal cell carcinoma staging with learnable image histogram-based deep neural network. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 533–540. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_61
18. Delahunt, B., Cheville, J.C., et al.: The international society of urological pathology (ISUP) grading system for renal cell carcinoma and other prognostic parameters. Am. J. Surg. Pathol. **37**, 1490–1504 (2013)
19. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks (2019)
20. Tian, K., et al.: Automated clear cell renal carcinoma grade classification with prognostic significance. PLoS ONE **14**(10), e0222641 (2019)
21. Yeh, F.-C., Parwani, A.V., Pantanowitz, L., Ho, C.: Automated grading of renal cell carcinoma using whole slide imaging. J. Pathol. Inform. **5**(1), 23 (2014)
22. Sun, X., et al.: Prediction of ISUP grading of clear cell renal cell carcinoma using support vector machine model based on CT images. Medicine **98**(14) (2019)
23. Cui, E., et al.: Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. Eur. Radiol. **30**, 2912–2921 (2020)
24. Zhao, Y., et al.: Deep learning based on MRI for differentiation of low- and high-grade in low-stage renal cell carcinoma. J. Magn. Reson. Imaging **52**(5), 1542–1549 (2020)
25. Heller, N., 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. 101821 (2020)
26. Heller, N., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes (2019)
27. Zhao, H., Li, H., Cheng, L.: Chapter 14 - data augmentation for medical image analysis. In: Burgos, N., Svoboda, D. (eds.) Biomedical Image Synthesis and Simulation. The MICCAI Society book Series, pp. 279–302. Academic Press (2022)
28. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
29. Shahinfar, S., Meek, P., Falzon, G.: "How many images do i need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Ecol. Inform. **57**, 101085 (2020)
30. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a python library for efficient loading, pre-processing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. **208**, 106236 (2021)
31. Akar, E., Kara, S., Akdemir, H., Kiriş, A.: Fractal analysis of MR images in patients with chiari malformation: the importance of pre-processing. Biomed. Signal Process. Control **31**, 63–70 (2017)
32. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual U-Net for medical image segmentation. J. Med. Imaging (Bellingham) **6**, 014006 (2019)
33. Vankdothu, R., Hameed, M.A.: Brain tumor MRI images identification and classification based on the recurrent convolutional neural network. Meas. Sens. 100412 (2022)
34. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
35. Tustison, N.J., et al.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**, 1310–1320 (2010)

36. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014)
37. Boone, L., et al.: ROOD-MRI: benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI (2022)
38. Krishna, S.T., Kalluri, H.K.: Deep learning and transfer learning approaches for image classification (2019)
39. Shaha, M., Pawar, M.: Transfer learning for image classification. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 656–660 (2018)
40. Hussain, M., Bird, J.J., Faria, D.R.: A study on CNN transfer learning for image classification. In: Lotfi, A., Bouchachia, H., Gegov, A., Langensiepen, C., McGinnity, M. (eds.) UKCI 2018. AISC, vol. 840, pp. 191–202. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-97982-3_16
41. Yang, A., Yang, X., Wu, W., Liu, H., Zhuansun, Y.: Research on feature extraction of tumor image based on convolutional neural network. IEEE Access **7**, 24204–24213 (2019)