Original Article

Non-annotated renal histopathological image analysis with deep ensemble learning

Jia Chun Koo[1]^, Qi Ke[2], Yan Chai Hum[1]^, Choon Hian Goh[1]^, Khin Wee Lai[3]^, Wun-She Yap[1]^, Yee Kai Tee[1]^

[1]Lee Kong Chian Faculty of Engineering and Science, University Tunku Abdul Rahman, Kajang, Malaysia; [2]School of Big Data and Artificial Intelligence, Guangxi University of Finance and Economics, Nanning, China; [3]Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia

Contributions: (I) Conception and design: JC Koo, Q Ke, CH Goh, YK Tee; (II) Administrative support: YC Hum, KW Lai, WS Yap, YK Tee; (III) Provision of study materials or patients: JC Koo, Q Ke, KW Lai, WS Yap; (IV) Collection and assembly of data: JC Koo, Q Ke, YC Hum; (V) Data analysis and interpretation: JC Koo, Q Ke, CH Goh, YK Tee; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Yee Kai Tee, DPhil. Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, Cheras 43000, Kajang, Selangor, Malaysia. Email: teeyeekai@gmail.com; teeyk@utar.edu.my.

Background: Renal cancer is one of the leading causes of cancer-related deaths worldwide, and early detection of renal cancer can significantly improve the patients' survival rate. However, the manual analysis of renal tissue in the current clinical practices is labor-intensive, prone to inter-pathologist variations and easy to miss the important cancer markers, especially in the early stage.

Methods: In this work, we developed deep convolutional neural network (CNN) based heterogeneous ensemble models for automated analysis of renal histopathological images without detailed annotations. The proposed method would first segment the histopathological tissue into patches with different magnification factors, then classify the generated patches into normal and tumor tissues using the pre-trained CNNs and lastly perform the deep ensemble learning to determine the final classification. The heterogeneous ensemble models consisted of CNN models from five deep learning architectures, namely VGG, ResNet, DenseNet, MobileNet, and EfficientNet. These CNN models were fine-tuned and used as base learners, they exhibited different performances and had great diversity in histopathological image analysis. The CNN models with superior classification accuracy (Acc) were then selected to undergo ensemble learning for the final classification. The performance of the investigated ensemble approaches was evaluated against the state-of-the-art literature.

Results: The performance evaluation demonstrated the superiority of the proposed best performing ensembled model: five-CNN based weighted averaging model, with an Acc (99%), specificity (Sp) (98%), F1-score (F1) (99%) and area under the receiver operating characteristic (ROC) curve (98%) but slightly inferior recall (Re) (99%) compared to the literature.

Conclusions: The outstanding robustness of the developed ensemble model with a superiorly high-performance scores in the evaluated metrics suggested its reliability as a diagnosis system for assisting the pathologists in analyzing the renal histopathological tissues. It is expected that the proposed ensemble deep CNN models can greatly improve the early detection of renal cancer by making the diagnosis process more efficient, and less misdetection and misdiagnosis; subsequently, leading to higher patients' survival rate.

Keywords: Deep learning; transfer learning; ensemble learning; histopathological image; renal cancer

^ ORCID: Jia Chun Koo, 0000-0002-3295-0583; Yan Chai Hum, 0000-0002-9657-8311; Choon Hian Goh, 0000-0002-8914-8524; Khin Wee Lai, 0000-0002-8602-0533; Wun-She Yap, 0000-0002-0007-6174; Yee Kai Tee, 0000-0002-0263-6358.

Submitted Jan 09, 2023. Accepted for publication Jul 03, 2023. Published online Jul 21, 2023.
doi: 10.21037/qims-23-46
View this article at: https://dx.doi.org/10.21037/qims-23-46

Introduction

Renal cancer, or also commonly known as kidney cancer, is one of the most common cancers, accounting for approximately 2% of global cancer diagnoses and deaths in 2020 (1). Two-third of the renal cancer patients are men, which makes renal cancer the seventh most frequently occurring cancer among men (1,2). Renal cell carcinoma (RCC) responsible for 85% of all renal cancer incidents, and clear cell renal cell carcinoma (CCRCC) originated from proximal tubule epithelial cells is the most common RCC subtype, contributing around 70% of the RCC cases (2,3). The survival rate of kidney cancer patients is highly dependent on the stage at diagnosis. Patients with clinically localized RCC (stage I) have 92.5% of 5-year survival rates, followed by 72.5% for regional RCC (stage II/III), and merely 12% for metastatic RCC (stage IV) (3). As a result, the early stage RCC diagnosis is important and decisive for effective patient treatment and thus saving lives.

Histopathological analysis is the clinical gold standard for malignancy diagnosis, subtype classification, or tumor grading on renal tissues. In the standard diagnostic and grading procedure of renal cancer, the pathologists have to manually evaluate the nucleoli and morphology of the renal tissues from low to high magnification power under a microscope (4-6). For instance, in renal tissues, basophilic tumor nucleoli are well captures in high-power magnification histopathological images, but for giant tumor cells such as eosinophilic tumor nucleoli are better captured at a lower magnification power (6). However, this conventional and manual visual analysis of renal tissues by pathologists is extremely laborious, time consuming, and subjective, where the conclusion drawn by a pathologist can be different from another. The correct analysis of renal tissues is highly dependent upon the experience and expertise of the pathologists. This makes the manual histopathological analysis prone to human errors such as misdetection and misdiagnosis, leading to a delay in the treatment and thus a lower survival rate (7).

The shortcomings of manual analysis arise the development of computer-aided diagnosis (CAD) systems to assist pathologists by providing an autonomous and efficient analysis of the histopathological images (8). The digitalization of biopsy slides and advancements in CAD systems have started to gain the attention of the medical community and enhance the overall analysis workflow. These systems not only reduce the time and cost of cancer diagnosis but also the inter-pathologist variability in diagnostic decisions (9). In this regard, various CAD systems have been developed, and recently, deep machine learning approaches are widely adopted in the systems for different analysis tasks due to its favorable characteristics and great performances (10).

The key components in deep learning techniques, convolutional neural networks (CNNs), have been widely used in CAD systems because the conventional algorithm-based machine learning approaches require supervision in feature extraction, whereas the use of deep CNNs can automatically learn high-level features without any domain knowledge (11-13). Many researchers have studied the performances of various CNN architectures at different model scale or depth. The studies demonstrated that although CNNs with greater scale or depth were able to result outstanding classification performance, they often have a significant drawback in execution time (9,14-16). Several lightweight architectures were proposed to address the drawback, which at the same time, able to yield a noticeable high accuracy (Acc) at a low latency (17-20).

Transfer learning has been frequently used in deep learning to train CNN models with limited dataset faster and accurately by transferring the knowledge obtained from a large-scale dataset to the target dataset for effective spatial features extraction. This technique is especially useful in the medical domain, where medical images and annotations are often more expensive and time-intensive to obtain (21,22). Fine-tuning is one of the transfer learning approaches which retrains several layers of the pretrained models to enable better feature extraction according to the target domain to obtain a better performance (23). Partially frozen of layers in CNN for fine-tuning was reported to have better classification performance compared to fully frozen or no frozen strategies as it managed better in both generic and specific features of the histopathological images (9,19,24).

Moreover, the complex nature of histopathological images makes it challenging to determine the distinguishing

features using a single classification approach. Therefore, ensemble learning techniques which involve multiple classification models are often implemented to deal with the complexity of histopathological images (9,25). Majority voting and averaging were the commonly used statistical approaches for improving the classification results from several base CNNs (9,26). Meanwhile, the use of neural network as classification layers for the combined base CNNs could perform slightly better than the statistical approaches (27).

The selection of image labelling method also plays an important role in histopathological image classification as it will gives different results and conclusions to the tissue samples. Patch-based labelling method is popular especially for histopathological images as it can perform cancer type discrimination, cancerous region indication, and prognosis analysis (27-32). However, such method not only requires expert pathologists to label the region-of-interest (ROI) accurately, but it is also leading back to the issues arose by manual histopathological image analysis, where the pathologists need to do this patch-by-patch and one whole slide image (WSI) can have thousands of patches depending on the image size and resolution (28,29). To alleviate the burden from high-detailed annotations, several studies have implemented WSI-based or slide-based labelling method and managed to yield very good results (28,33-35). The use of multiple instance learning (MIL) approach can further extend the capability of WSI-level labelling to classify the patches more accurately without the need of patch-level annotations (36,37). Yet, there are limited studies on the feasibility of non-ROI-annotation classification, particularly when it comes to different magnification factors of the histopathological images.

This work aims to investigate the performance of heterogeneous ensembled deep learning models in the diagnosis of annotation-free and multi-magnification renal histopathological images. For this purpose, tissue segmentation and extraction of WSIs into smaller image patches were done initially for the image pre-processing steps. Five different pre-trained CNN architectures and their variants were used to discriminate the different magnification factors renal image patches into normal and tumor classes respectively. The best performed deep learning models were then combined with different ensemble techniques to classify the renal histopathological images. Lastly, the studied models were evaluated to determine their reliability as a good diagnosis system. In this regards, early diagnosis of renal cancer can be done with the best investigated model, to prevent its progression and lower its morbidity rates in renal cancer patients. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-46/rc).

Methods

The WSIs with extremely high image resolutions underwent image pre-processing steps to generate image patches with smaller dimensions before they were fed to the deep learning models. The image patches generated in different magnification factors were then used for model training, validation, and testing by using the pre-trained CNNs with transfer learning. The best three and five models with the highest Acc underwent ensemble learning to perform the final classification. The proposed models were evaluated with various metrics. The overall workflow of this study is depicted in *Figure 1*.

Dataset preparation

The Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma (CPTAC-CCRCC) dataset (38) from The Cancer Imaging Archive (TCIA) was used in this study to evaluate the performance of the proposed method in classifying the renal histopathological images into normal and tumor classes. The dataset consists of 783 hematoxylin and eosin (H&E) stained WSIs, composed by 259 normal and 524 tumor (particularly CCRCC) renal tissue slide images. The WSIs were scanned by Aperio digital pathology slide scanner manufactured by Leica, at a 40× magnification in red, green, and blue (RGB) color system, and the slides are consistently having a (0.494, 0.494) mpp pixel aspect ratio (PAR). Subtypes of CCRCC: rhabdoid variant CCRCC and partly papillary CCRCC, in the dataset are grouped together, all these were classified as the tumor class. In the tumor WSIs, all the tissues were assumed to be tumor although it was possible to have some normal or non-tumor area in the slide since detailed annotations of tumor lesions were not provided. A stratified hold-out validation method was used for dataset splitting to ensure each set has similar distribution of classes. The dataset was partitioned into 70% (181 normal and 366 tumor WSIs), 20% (52 normal and 105 tumor WSIs), and 10% (26 normal and 53 tumor WSIs) for the training, validation, and testing sets, respectively. Although the dataset is imbalanced, where the tumor class has contributed approximately 67% in the

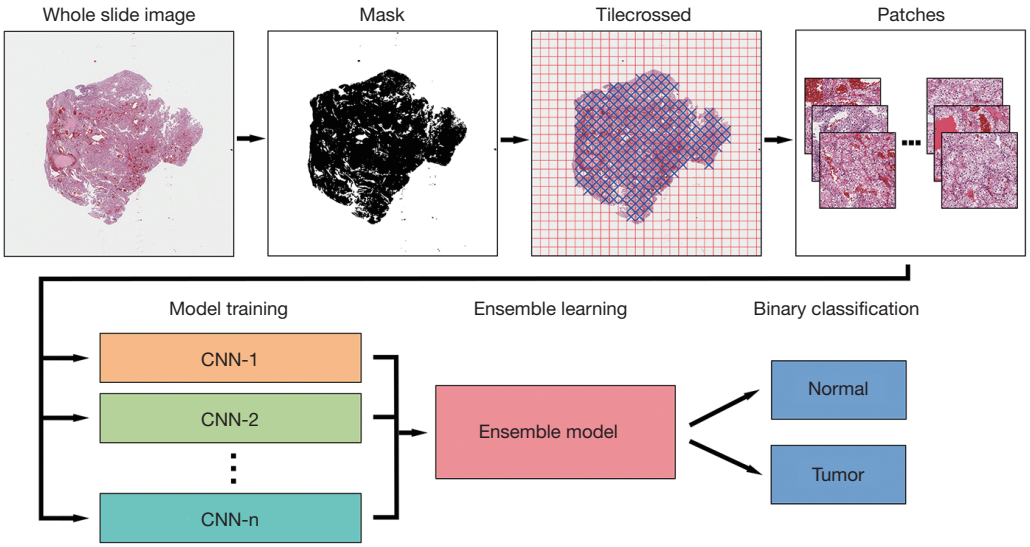

Figure 1 Overall schematic workflow. The H&E stained WSI is tiled into magnified (e.g., 2.5× magnification factor) image patches and fed to the CNN models, followed by ensemble model, for classification into normal or tumor class. H&E, hematoxylin and eosin; WSI, whole slide image; CNN, convolutional neural network.

dataset, dataset imbalance treatment was not performed. This is because a greater contribution in tumor class reflects the actual scenario in histopathological studies; hence, maintaining the class distribution can make the proposed classification model more apt to actual histopathological situation. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Histopathological image pre-processing

PyHIST (39) histopathological image pre-processing tool proposed by Muñoz-Aguirre *et al.* was used to generate image patches from the WSIs. The pipeline of PyHIST consists of three main processing steps: tissue masking, tiling, and patch extraction. A mask was generated with Otsu thresholding method for the histopathological image, followed by the generation of a grid of non-overlapping tiles on top of the mask with a tile size of 224×224 pixels. The output downsampling factors selected for the patches were 16, 8, 4, and 2, which representing magnification levels of 1.25×, 2.5×, 5×, and 10×, respectively, as shown in *Figure 2*. Meanwhile, both mask downsampling and tile crossed image downsampling were set as defaults.

The content threshold was set to 0.5 to extract tiles with 50% or more tissue coverage. Although image patches with greater coverage of tissues is favorable as they have less non-informative area, the number of generated patches needs to be considered. Lesser image patches can be generated when the threshold is set to be higher. However, this can lead to insufficient histopathological images for model training, which can be significant especially for the low magnification setting. One of the major drawbacks of PyHIST is that it does not remove artefacts such as blur, folded tissue, and pen ink from the WSI before masking, which reduces the quality of histopathological image patches for classification. Thus, manual image data cleaning was performed to remove undesirable image patches generated. The number of image patches after data cleaning for each magnification factors is tabulated in *Table 1*. After image patches were generated from the WSIs, each of the patches was labelled according to its corresponding WSI-level label without ROI.

Classification model development

In this work, five CNN architectures (VGG, ResNet, DenseNet, MobileNet, and EfficientNet) and their variants were applied for transfer learning with pretrained weights from ImageNet to classify the histopathological images of renal cancer in the CPTAC-CCRCC dataset. Fine-tuning technique was adopted for transfer leaning and its schematic diagram is shown in *Figure 3*. In the fine-tuning approach, 80% of the pretrained layers in convolutional base were frozen while another 20% of the layers were remain unfrozen for model re-training with the selected

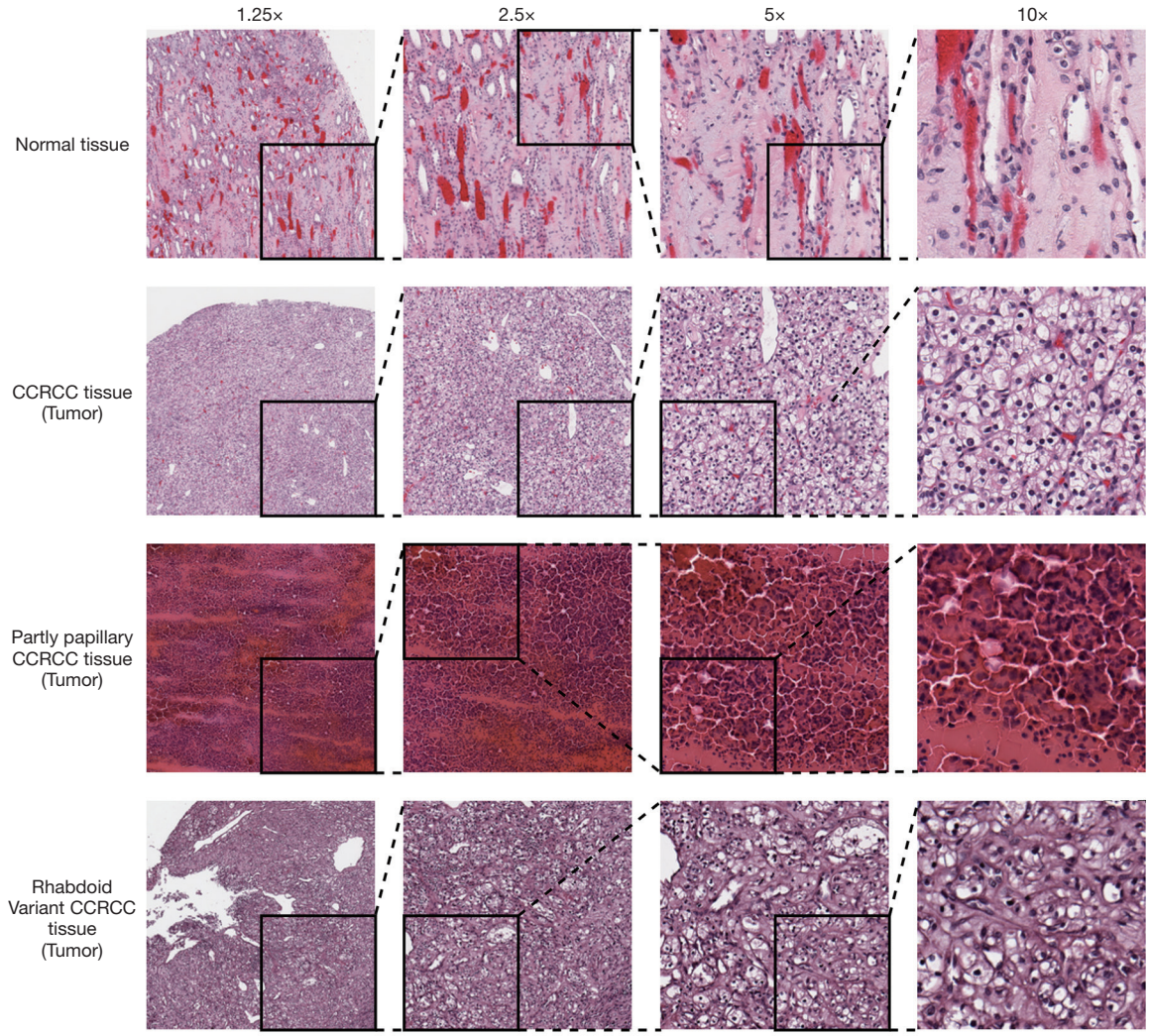

Figure 2 Sample generated H&E stained normal and CCRCC tissue images patches with four different magnification factors. H&E, hematoxylin and eosin; CCRCC, clear cell renal cell carcinoma.

Table 1 Number of image patches after data cleaning for each magnification factor

Dataset	WSI	Magnification factor			
		1.25×	2.5×	5×	10×
Training	547	12,803	52,479	211,576	843,986
Validation	157	3,507	14,426	58,055	231,501
Testing	79	1,936	7,945	32,050	127,629
Total	783	18,246	74,850	301,681	1,203,116

WSI, whole slide image.

histopathological dataset.

Four new trainable classification layers were created to replace the excluded pretrained classification layers for histopathological image classification. A global average pooling layer was appended to flatten the vectors into single dimension, followed by a fully connected layer with 512 neurons and a rectified linear unit (ReLU) activation function. A dropout layer with a rate of 0.5 was added to further minimize overfitting and lastly, a fully connected sigmoid layer was added as a classifier to perform the binary

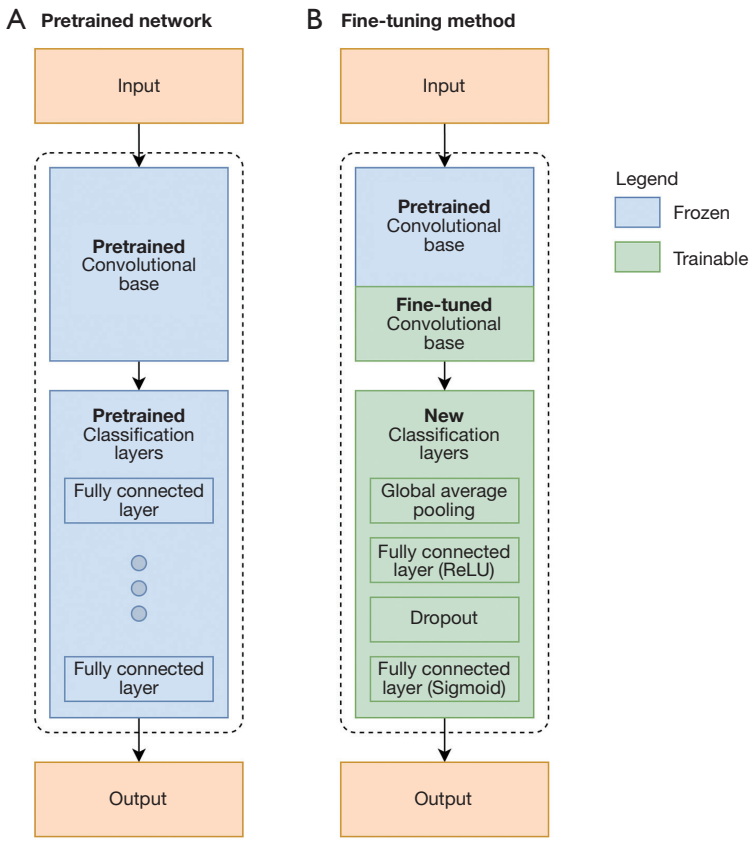

Figure 3 An overview of transfer learning technique. (A) Pretrained network; (B) fine-tuned network. ReLU, rectified linear unit.

classification on the histopathological images. AdaBound (40) optimizer which inherited the advantages of Adam and stochastic gradient descent (SGD) optimizers was selected to optimize the new layers on the histopathological images with a very low learning rate of 0.0001, and binary cross-entropy was used as the loss function.

In the model training, an input image size of [224, 224, 3] was used for all the CNN models. The models were trained at different magnification factors separately with a maximum of 50 epochs and a batch size of 32. An early-stopping callback with a patience of 5 was implemented to obtain the epoch with the best validation Acc effectively and mitigate the possibility of overfitting. The best weight of the model was saved after the completion of training and the models with a high Acc would undergo ensemble learning to perform the final classification.

Ensemble learning

It is clearly proven in the literature (9,26,27) that the adoption of ensemble learning is able to improve the generalization of the deep learning system and leads to a better prediction. Ensemble leaning combines the decision of multiple classifiers, either at final or intermediate stages, to overcome the limited capacity of a single classifier and hence improve the robustness of the classification framework. A wide variety of ensemble techniques are used to develop a robust and accurate ensemble model, such as voting, bagging, boosting, and stacking.

In this work, four multi-model ensemble approaches, namely majority voting, unweighted averaging, weighted averaging, and stacking, were used to yield a more accurate and robust classification performance. The number of models in ensemble was often kept small to avoid the diminishing returns in performance. Therefore, the CNN models with best performance, particularly ones with the top-3 and top-5 highest Acc, were selected for ensemble learning.

Majority voting

Ensemble learning based on majority voting (*Figure 4A*),

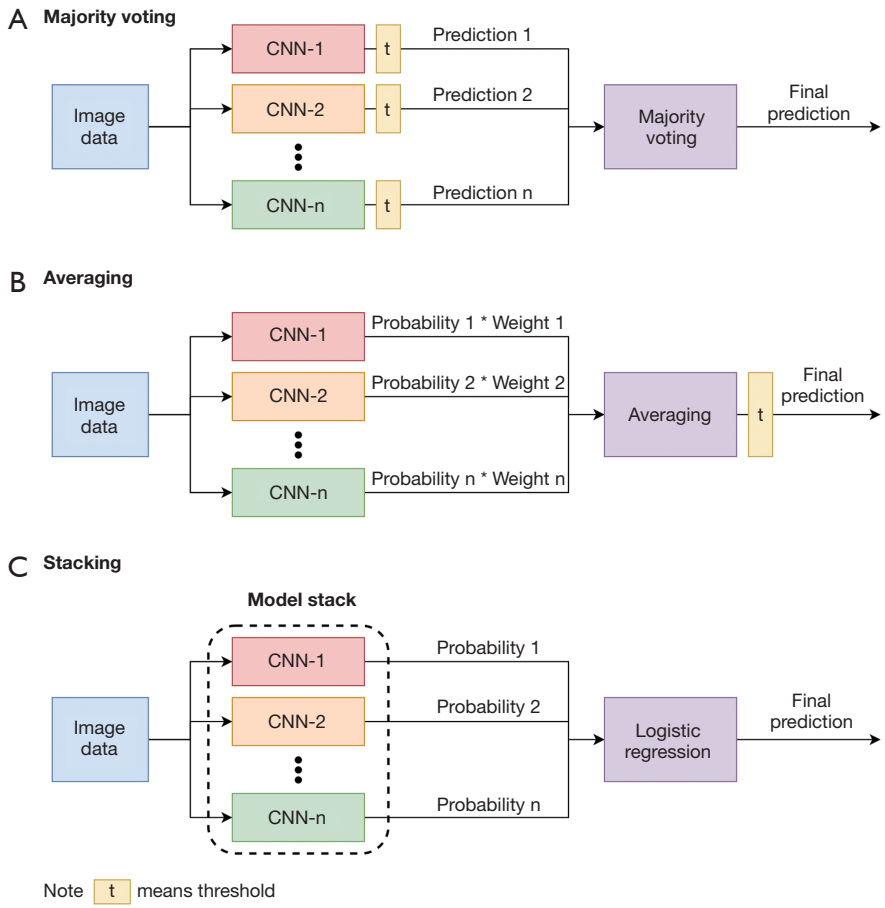

Figure 4 An overview of ensemble learning techniques. (A) Majority voting; (B) averaging; (C) stacking. CNN, convolutional neural network.

or so-called hard voting, takes the mode of the predicted class from each of the CNN models as final prediction. The number of models involved in ensemble learning, including majority voting, was set to odd number to avoid the scenario of having two modes during voting. Since each of the individual CNN models stands an equal chance for voting the final prediction class, it will be less biased towards the prediction of a particular CNN model as the effect is mitigated by the majority vote count. However, it is possible to result a dominance of event when majority of the similar base learners are having identical preference of a particular event (41). Hence, majority voting will perform better when CNN models from different architectures are involved for ensemble learning.

Averaging

Averaging-ensemble learning (*Figure 4B*) is another widely used technique, and both unweighted and weighted averaging were implemented in this study. Both averaging approaches took the mean of the probability predicted from each model, and then predicted the class of the histopathology image based on the new probabilistic score resulted from the ensemble model. In unweighted averaging ensemble, each classifier was designed to have equal contribution to the final probability score. In contrast, for the weighted averaging ensemble, each of the contributing models had different weightage and the weight coefficients were manipulated empirically until the best result was obtained. It has advantage over the unweighted averaging especially when the models have considerable different performance, where model with better result can have a greater weight coefficient. Meanwhile, unweighted averaging is preferred when the models share similar performance (42).

Stacking

Stacking (*Figure 4C*) is an ensemble learning approach which uses a meta-learner to integrate the output of multiple base classifiers. There are usually two levels of classifiers in stacking. Stacking starts by training the heterogenous base classifiers at the first level, which are the selected CNNs in this case, to generate a new dataset for the meta-learner in second level. There would be a chance of overfitting if the new generated training set consists of exact training samples used by the base learners at first level. Hence, validation set is employed for the base classifiers to generate a clean training set for the meta-learner in second level. Currently, various machine learning algorithms have been used as a meta-learner, such as Naïve Bayes, random forests, and gradient boosting machine (43). Logistic regression was selected as a meta-learner in this study as it is the most popular binary classification scheme used in stacking.

Model evaluation and visualization

Evaluation metrics

The performance of the trained CNN models and ensembled models was evaluated by using evaluation metrics: Acc, specificity (Sp), precision (Pr), recall (Re), and area under the curve (AUC) of receiver operating characteristic (ROC) curve. Since the selected dataset has a class imbalance problem, the F1-score (F1) was also computed as it is an unbiased estimator which provides equal weightage to Pr and Re. Moreover, the Matthews correlation coefficient (*MCC*) was used as one of the evaluation criteria. It is a useful measure in imbalanced dataset as it considered all the confusion metrics: true positive (*TP*), false negative (*FN*), true negative (*TN*), and false positive (*FP*), proportionally both to the positive and negative classes. Besides, the reliability of the proposed deep learning model as a diagnosis system was also considered. Xie *et al.* (11) reported that a reliable diagnosis system should achieve the requirement of Re ≥80%, Sp ≥95%, Pr ≥95%, and diagnostic odds ratio (*DOR*) ≥100. Hence, the proposed methods were evaluated according to the stated requirements above to identify the fulfilment as a reliable diagnosis system. The mathematical formulations of *MCC* and *DOR* are presented in Eqs. [1,2]:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad [1]$$

$$DOR = \frac{TP \times TN}{FP \times FN} \quad [2]$$

Prediction visualization

Gradient weighted class activation mapping (Grad-CAM) technique was implemented to visualize how the models predict and classify the renal histopathological images into normal and tumor class. The feature maps from the final convolutional layer of the model were used in Grad-CAM to create an activated heatmap, which then superimposed on the histopathological image patches to highlight the visual patterns for the class prediction. This facilitated the understanding of the predictions from the 'black-box' of the deep learning models. Since the CNN models from the same architecture tend to have a similar visual pattern, a single CNN model would be selected from each of the three top-performing architectures for prediction visualization, to study the heatmap diversity from different architectures.

Experimental setup

Python (version 3.8.8) and TensorFlow (version 2.3.0) framework were used in Jupyter Notebook to run all the experiments in this study. The image pre-processing and models training were done on a Windows 10 workstation which equipped with dual Intel XEON E5-2630cv3 processors, NVIDIA Quadro P6000 GPU, 120 GB of RAM, and CUDA version of 11.4.

Results

In this work, CNN models from five architectures (VGG, ResNet, DenseNet, MobileNet, and EfficientNet) and four ensemble techniques were used to classify the renal histopathological images. The results for 1.25× (*Table 2*), 2.5× (*Table 3*), 5× (*Table 4*), and 10× (*Table 5*) magnification factors on testing set are shown below. The ensembled model names are abbreviated in the tables as "MV", "UA", "WA", and "ST" for majority voting, unweighted averaging, weighted averaging, and stacking, respectively. Meanwhile, the annotation behind the abbreviated ensembled models stand for the number of individual CNN models involved for ensemble learning. For instance, "Ensemble-MV3" means the model ensembles the top three highest-Acc CNN models with majority voting technique. The performances of the investigated individual CNN models and ensembled models are evaluated based on the evaluation metrics mentioned in section "Evaluation metrics". The detailed results are discussed in the following subsections.

Table 2 Classification results of fine-tuned and ensembled models for 1.25× magnification factor

Architecture	Acc	Sp	Pr	Re	F1	AUC	MCC	DOR
VGG16	0.9685	0.9692	0.9854	0.9682	0.9767	0.9687	0.9284	956
VGG19	0.9706	0.9286	0.9674	0.9902	0.9787	0.9594	0.9319	1,307
ResNet50	0.9659	0.9383	0.9714	0.9788	0.9751	0.9585	0.9212	702
ResNet101	0.9742	0.9351	0.9704	0.9924	0.9813	0.9637	0.9403	1,886
ResNet152	0.9669	0.9724	0.9868	0.9644	0.9755	0.9684	0.9254	954
DenseNet121	0.9478	0.8620	0.9388	0.9879	0.9627	0.9249	0.8792	509
DenseNet169	0.9711	0.9643	0.9832	0.9742	0.9787	0.9693	0.9338	1,021
DenseNet201	0.9504	0.8864	0.9487	0.9803	0.9642	0.9333	0.8848	388
MobileNet	0.8760	0.7565	0.8913	0.9318	0.9111	0.8442	0.7086	42
MobileNetV2	0.7025	0.6964	0.8327	0.7053	0.7637	0.7009	0.3788	5
EfficientNetB0	0.9742	0.9513	0.9774	0.9848	0.9811	0.9681	0.9403	1,270
EfficientNetB1	0.9726	0.9756	0.9884	0.9712	0.9797	0.9734	0.9379	1,352
EfficientNetB2	0.9762	0.9708	0.9863	0.9788	0.9825	0.9748	0.9455	1,533
EfficientNetB3	0.9659	0.9594	0.9808	0.9689	0.9748	0.9642	0.9221	737
EfficientNetB4	0.9602	0.9383	0.9712	0.9705	0.9708	0.9544	0.9084	500
Ensemble-MV3	0.9819	0.9610	0.9820	0.9917	0.9868	0.9764	0.9582	2,935
Ensemble-UA3	0.9809	0.9594	0.9812	0.9909	0.9861	0.9752	0.9558	2,577
Ensemble-WA3	0.9809	0.9497	0.9770	**0.9955**	0.9861	0.9726	0.9559	4,133
Ensemble-ST3	0.9809	0.9675	0.9849	0.9871	0.9860	0.9773	0.9559	2,284
Ensemble-MV5	**0.9861**	**0.9756**	**0.9887**	0.9909	**0.9898**	**0.9833**	**0.9678**	**4,367**
Ensemble-UA5	0.9788	0.9545	0.9790	0.9902	0.9846	0.9723	0.9510	2,111
Ensemble-WA5	0.9845	0.9659	0.9842	0.9932	0.9887	0.9795	0.9642	4,127
Ensemble-ST5	0.9824	0.9724	0.9871	0.9871	0.9871	0.9798	0.9595	2,701

The best results of the metrics are bolded; and 3 or 5 after the ensemble model abbreviation indicates top 3 or top 5 CNN models. Acc, accuracy; Sp, specificity; Pr, precision; Re, recall; F1, F1-score; AUC, area under the curve; MCC, Matthews correlation coefficient; DOR, diagnostic odds ratio; MV, majority voting; UA, unweighted averaging; WA, weighted averaging; ST, stacking; CNN, convolutional neural network.

Performance analysis of CNNs

For the VGG architecture, VGG16 variant acquired a higher Acc by at least 0.3% compared to VGG19 in most of the settings, except for 1.25× magnification factor case where VGG19 yielded an Acc 0.2% higher than VGG16. In term of Sp, VGG16 had superior performance than VGG19, meaning that VGG16 can raise less FPs than VGG19. Since a reliable diagnosis system should have equal or more than 95% of Sp, VGG19 was not a good candidate as it only fulfilled the requirement at 2.5× magnification factor. Besides, both VGG variants had different behavior in Pr and Re, where VGG16 tends to have better Pr while VGG19 is better in achieving high Re in most of the magnification factors. Even though VGG16 has shallower layers than VGG19, it managed to perform better than VGG19 in various evaluated metrics except Re.

In the ResNet architecture, both ResNet101 and ResNet152 had great performance in classifying the histopathological images with different magnification factors. However, these two models exhibited different strengths, such that ResNet101 had higher Acc and F1,

Table 3 Classification results of fine-tuned and ensembled models for 2.5× magnification factor

Architecture	Acc	Sp	Pr	Re	F1	AUC	MCC	DOR
VGG16	0.9790	0.9759	0.9889	0.9804	0.9846	0.9782	0.9515	2,026
VGG19	0.9753	0.9603	0.9818	0.9822	0.9820	0.9712	0.9427	1,335
ResNet50	0.9727	0.9671	0.9848	0.9752	0.9800	0.9712	0.9371	1,158
ResNet101	0.9763	0.9663	0.9845	0.9809	0.9827	0.9736	0.9452	1,475
ResNet152	0.9773	0.9687	0.9856	0.9813	0.9835	0.9750	0.9476	1,624
DenseNet121	0.9524	0.9274	0.9667	0.9639	0.9653	0.9456	0.8897	341
DenseNet169	0.9572	0.9310	0.9685	0.9692	0.9688	0.9501	0.9006	424
DenseNet201	0.9590	0.9402	0.9725	0.9675	0.9700	0.9539	0.9050	469
MobileNet	0.9091	0.8404	0.9280	0.9406	0.9342	0.8905	0.7875	83
MobileNetV2	0.8682	0.9021	0.9501	0.8527	0.8988	0.8774	0.7201	53
EfficientNetB0	0.9787	0.9655	0.9842	0.9848	0.9845	0.9751	0.9506	1,810
EfficientNetB1	0.9806	0.9811	0.9913	0.9804	0.9858	0.9808	0.9554	2,600
EfficientNetB2	0.9767	0.9775	0.9896	0.9763	0.9829	0.9769	0.9466	1,796
EfficientNetB3	0.9806	0.9739	0.9880	0.9837	0.9858	0.9788	0.9551	2,251
EfficientNetB4	0.9766	0.9691	0.9858	0.9800	0.9829	0.9746	0.9459	1,538
Ensemble-MV3	0.9851	0.9824	0.9919	0.9864	0.9891	0.9844	0.9657	4,045
Ensemble-UA3	0.9850	0.9828	0.9921	0.9861	0.9891	0.9844	0.9654	4,030
Ensemble-WA3	0.9850	0.9795	0.9906	**0.9875**	0.9891	0.9835	0.9653	3,791
Ensemble-ST3	0.9845	0.9819	0.9917	0.9857	0.9887	0.9838	0.9642	3,748
Ensemble-MV5	0.9846	0.9811	0.9913	0.9862	0.9888	0.9837	0.9645	3,731
Ensemble-UA5	0.9814	0.9795	0.9906	0.9822	0.9864	0.9809	0.9570	2,643
Ensemble-WA5	**0.9855**	0.9819	0.9917	0.9872	**0.9894**	**0.9846**	**0.9665**	**4,183**
Ensemble-ST5	0.9850	**0.9832**	**0.9922**	0.9859	0.9891	0.9845	0.9654	4,074

The best results of the metrics are bolded; and 3 or 5 after the ensemble model abbreviation indicates top 3 or top 5 CNN models. Acc, accuracy; Sp, specificity; Pr, precision; Re, recall; F1, F1-score; AUC, area under the curve; MCC, Matthews correlation coefficient; DOR, diagnostic odds ratio; MV, majority voting; UA, unweighted averaging; WA, weighted averaging; ST, stacking; CNN, convolutional neural network.

which means it has better balance between Pr and Re than ResNet152. In contrast, ResNet152 had better Sp, Pr, and AUC, giving a better capability in distinguishing between normal and tumor classes. Meanwhile, ResNet50 with lesser layers did not surpass both ResNet101 and ResNet152 in any of the evaluated components, and its Sp did not meet the requirement as a reliable diagnosis system when the magnification factor was set to be very high (10×) or low (1.25×).

The DenseNet architecture had less satisfactory results compared to VGG and ResNet, where the accuracies were mostly below 96%. Overall, DenseNet201 was the best variant as it achieved the highest results in most of the evaluated criteria. But when it comes to images with lower magnification factor (1.25×), DenseNet169 had better performance than DenseNet201. Similar to the ResNet architecture, DenseNet variants with deeper layers achieved better classification performance in renal histopathological images. However, all of the variants, except DenseNet169 for 1.25× magnification factor, yielded Sp less than 95%, which means DenseNet can hardly be a reliable diagnosis system with the setting in this study. This could be due

Table 4 Classification results of fine-tuned and ensembled models for 5× magnification factor

Architecture	Acc	Sp	Pr	Re	F1	AUC	MCC	DOR
VGG16	0.9723	0.9605	0.9818	0.9777	0.9798	0.9691	0.9360	1,068
VGG19	0.9673	0.9358	0.9709	0.9818	0.9763	0.9588	0.9238	784
ResNet50	0.9682	0.9610	0.9819	0.9715	0.9767	0.9662	0.9268	839
ResNet101	0.9726	0.9585	0.9809	0.9790	0.9800	0.9688	0.9365	1,078
ResNet152	0.9717	0.9654	0.9840	0.9746	0.9793	0.9700	0.9348	1,070
DenseNet121	0.9521	0.9119	0.9600	0.9706	0.9653	0.9412	0.8883	341
DenseNet169	0.9587	0.9366	0.9709	0.9689	0.9699	0.9527	0.9044	460
DenseNet201	0.9603	0.9402	0.9725	0.9696	0.9710	0.9549	0.9082	501
MobileNet	0.9352	0.8846	0.9477	0.9584	0.9530	0.9215	0.8487	176
MobileNetV2	0.8844	0.6827	0.8704	0.9769	0.9206	0.8298	0.7272	91
EfficientNetB0	0.9783	0.9732	0.9876	0.9806	0.9841	0.9769	0.9499	1,837
EfficientNetB1	0.9802	0.9774	0.9895	0.9816	0.9855	0.9795	0.9545	2,300
EfficientNetB2	0.9757	0.9661	0.9844	0.9801	0.9822	0.9731	0.9437	1,400
EfficientNetB3	0.9752	0.9649	0.9838	0.9800	0.9819	0.9724	0.9427	1,344
EfficientNetB4	0.9756	0.9614	0.9823	0.9821	0.9822	0.9718	0.9434	1,368
Ensemble-MV3	0.9820	**0.9781**	**0.9899**	0.9838	0.9869	0.9810	0.9585	2,716
Ensemble-UA3	0.9821	0.9779	0.9898	0.9840	0.9869	0.9809	0.9586	2,722
Ensemble-WA3	0.9825	0.9766	0.9892	**0.9852**	0.9872	0.9809	0.9595	2,778
Ensemble-ST3	0.9820	0.9780	0.9898	0.9839	0.9869	0.9809	0.9585	2,711
Ensemble-MV5	0.9826	0.9769	0.9894	**0.9852**	**0.9873**	0.9810	0.9597	2,814
Ensemble-UA5	0.9799	0.9734	0.9877	0.9828	0.9853	0.9781	0.9535	2,097
Ensemble-WA5	**0.9827**	0.9775	0.9896	0.9851	**0.9873**	**0.9813**	**0.9600**	**2,863**
Ensemble-ST5	0.9823	0.9768	0.9893	0.9848	0.9871	0.9808	0.9591	2,734

The best results of the metrics are bolded; and 3 or 5 after the ensemble model abbreviation indicates top 3 or top 5 CNN models. Acc, accuracy; Sp, specificity; Pr, precision; Re, recall; F1, F1-score; AUC, area under the curve; MCC, Matthews correlation coefficient; DOR, diagnostic odds ratio; MV, majority voting; UA, unweighted averaging; WA, weighted averaging; ST, stacking; CNN, convolutional neural network.

to the easily-overfitting characteristic of the DenseNet, leading a high FP rate in the imbalanced dataset.

The MobileNet architecture had the least favorable results among the selected CNN architectures, where both of the MobileNet variants had most of the performance metrics scored below the average. Moreover, none of the variants fulfilled the requirements as a reliable diagnosis system in any of the magnification factors. Since there is an incremental trend in performances when the magnification factor increases, and number of images generated are dependent on the magnification factor, the poor performances could be due to the insufficient trainable images for MobileNets, which limited the training quality of MobileNet models with the settings in this study. It was notable that MobileNet outperformed MobileNetV2 in every component of the performance metrics, which could be due to its parameter count greater than MobileNetV2. Considering the consequences of false classification in renal histopathology images, which could reduce cancer patient survival rate, MobileNet models with poor performances in the setting are less desirable.

The EfficientNet architecture was the best out of the five

Table 5 Classification results of fine-tuned and ensembled models for 10× magnification factor

Architecture	Acc	Sp	Pr	Re	F1	AUC	MCC	DOR
VGG16	0.9671	0.9428	0.9737	0.9783	0.9760	0.9605	0.9236	742
VGG19	0.9637	0.9342	0.9699	0.9773	0.9735	0.9558	0.9157	611
ResNet50	0.9690	0.9485	0.9763	0.9785	0.9774	0.9635	0.9282	837
ResNet101	0.9701	0.9534	0.9785	0.9779	0.9782	0.9657	0.9310	905
ResNet152	0.9684	0.9400	0.9725	0.9816	0.9770	0.9608	0.9267	834
DenseNet121	0.9526	0.9122	0.9599	0.9712	0.9656	0.9417	0.8897	351
DenseNet169	0.9522	0.9189	0.9627	0.9676	0.9652	0.9432	0.8891	338
DenseNet201	0.9590	0.9192	0.9632	0.9774	0.9702	0.9483	0.9046	491
MobileNet	0.9215	0.8501	0.9323	0.9544	0.9433	0.9022	0.8164	119
MobileNetV2	0.8717	0.7628	0.8937	0.9220	0.9077	0.8424	0.6985	38
EfficientNetB0	0.9699	0.9571	0.9801	0.9759	0.9780	0.9665	0.9307	902
EfficientNetB1	0.9688	0.9483	0.9762	0.9783	0.9772	0.9633	0.9278	827
EfficientNetB2	0.9736	0.9590	0.9810	0.9804	0.9807	0.9697	0.9390	1,170
EfficientNetB3	0.9677	0.9518	0.9777	0.9750	0.9763	0.9634	0.9253	770
EfficientNetB4	0.9713	0.9528	0.9782	0.9798	0.9790	0.9663	0.9335	980
Ensemble-MV3	0.9764	0.9618	0.9824	0.9831	0.9827	0.9724	0.9453	1,464
Ensemble-UA3	0.9767	0.9625	0.9827	0.9833	0.9830	0.9729	0.9461	1,506
Ensemble-WA3	0.9769	0.9608	0.9819	**0.9844**	0.9831	0.9726	0.9465	1,540
Ensemble-ST3	0.9765	0.9621	0.9825	0.9832	0.9828	0.9726	0.9456	1,483
Ensemble-MV5	0.9769	0.9631	0.9830	0.9833	0.9831	0.9732	0.9466	1,540
Ensemble-UA5	0.9727	0.9566	0.9800	0.9801	0.9800	0.9683	0.9368	1,085
Ensemble-WA5	**0.9777**	**0.9649**	**0.9838**	0.9836	**0.9837**	**0.9742**	**0.9483**	**1,645**
Ensemble-ST5	0.9773	0.9643	0.9835	0.9833	0.9834	0.9738	0.9476	1,596

The best results of the metrics are bolded; and 3 or 5 after the ensemble model abbreviation indicates top 3 or top 5 CNN models. Acc, accuracy; Sp, specificity; Pr, precision; Re, recall; F1, F1-score; AUC, area under the curve; MCC, Matthews correlation coefficient; DOR, diagnostic odds ratio; MV, majority voting; UA, unweighted averaging; WA, weighted averaging; ST, stacking; CNN, convolutional neural network.

selected architectures as it yielded the most models with the highest Acc in all magnification factors. EfficientNetB1 and B2 showed most outstanding performances in various evaluated metrics; hence, small to moderate compound scaling of parameters in EfficientNet was more suitable for this study setting. Interestingly, none one of the selected EfficientNet models scored below 90% for any of the performance metrics, and only one model (EfficientNetB4) from 1.25× magnification factor failed to fulfil the Sp criteria as a reliable diagnosis system. Therefore, EfficientNet was considered the best candidate for renal histopathological image classification.

Diversity analysis of CNN architectures

The diversity of the CNN architectures is evaluated by analyzing the variations of their detection responses. Three CNN models from different architectures: VGG16, ResNet101, and EfficientNetB0, were selected as the candidates for Grad-CAM heatmap generation on normal and tumor classes with four magnification factors. *Figure 5* shows the original image patches and the correctly classified

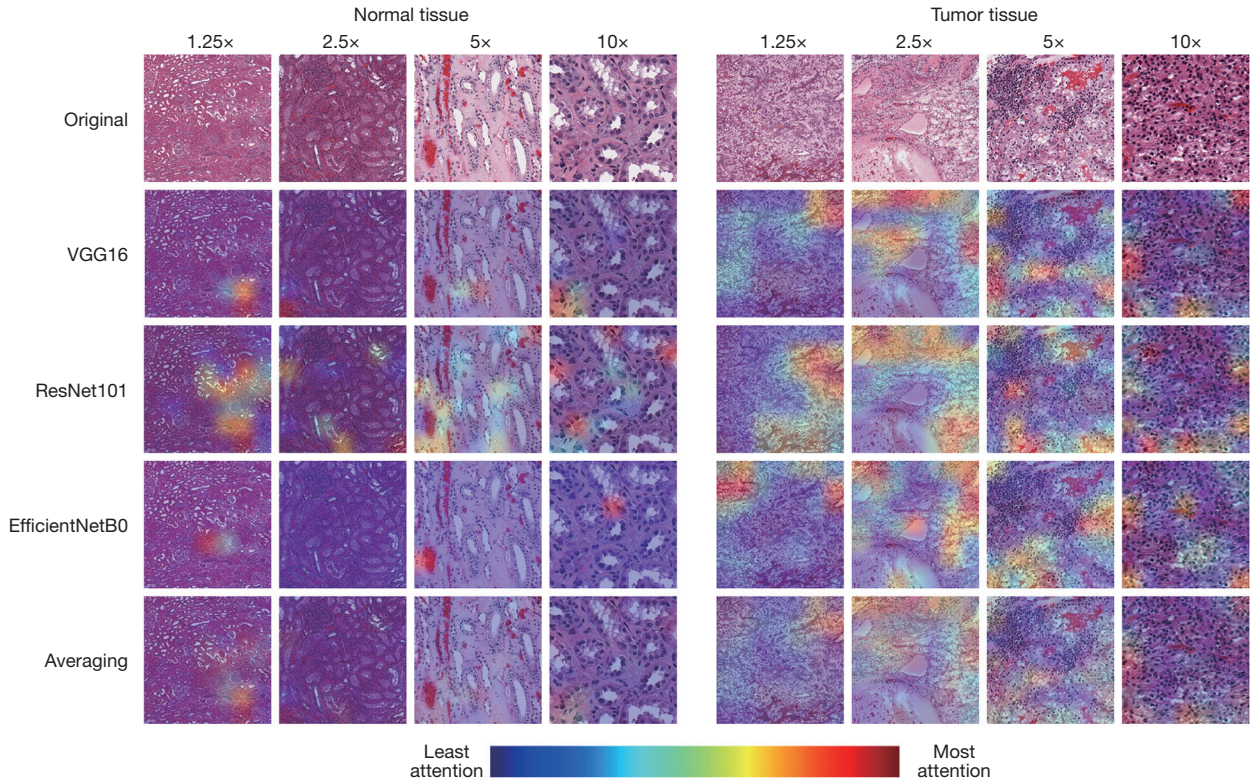

Figure 5 Three sample individual CNNs from different architectures response towards H&E stained normal and tumor histopathological image patches with different magnification factors. CNN, convolutional neural network; H&E, hematoxylin and eosin.

activated heatmaps produced by the three CNN models. The fluctuation in the heatmap intensity value correlates to the degree of attention given by the model on the highlighted locations for making decision. For instance, the region with warmer color (either red or near-red color) contributes the most attention by the model for the class prediction.

The activated areas of VGG16 and EfficientNetB0 are highly similar for normal tissues in various magnification factors, except that EfficientNetB0 tends to localize relatively smaller regions for decision making. Meanwhile, the heatmaps of ResNet101 are highly activated over the entire normal tissue image compared to VGG16 and EfficientNetB0, and this could be difficult for the pathologists to interpret the images as normal tissues due to the great coverage of high activated areas. For tumor tissues, there are no significant similarity in activated regions for three of the models. For instance, for tumor tissue with 1.25× magnification factor in *Figure 5*, VGG16 only highlighted the upper right region of the image while ResNet101 and EfficientNetB0 have additional highlighted areas at bottom and upper left, respectively. However, these three of the models still managed to recognize these regions as tumor tissues, which demonstrated the models' logic and capability in distinguishing normal and tumor tissues. The high variation in detection response between architectures is favorable especially in ensemble learning, where different areas will be considered for more robust and accurate predictions.

Performance analysis of ensembled models

Several CNNs based heterogeneous ensemble techniques had been deployed to classify the renal histopathological images effectively. Majority voting was the first ensemble technique tested in this study, and it shows a relatively promising results especially with low magnification factor (1.25×) or small amount of image patches. Increasing the number of individual CNN models to five is possible to improve the Acc at most up to 0.41% (1.25× magnification factor case) compared to majority voting with three CNN models. Surprisingly, there was no significant robustness

Table 6 Top-5 best performing models with their weightage used in weighted averaging ensemble learning method

Ranking	Magnification factor			
	1.25×	2.5×	5×	10×
1st	EfficientNetB2 (1.4)	EfficientNetB1 (1.4)	EfficientNetB1 (1.4)	EfficientNetB2 (1.4)
2nd	ResNet101 (1.2)	EfficientNetB3 (1.1)	EfficientNetB0 (0.9)	EfficientNetB4 (0.9)
3rd	EfficientNetB0 (1.2)	VGG16 (1.1)	EfficientNetB2 (0.9)	ResNet101 (0.9)
4th	EfficientNetB1 (1.05)	EfficientNetB0 (0.85)	EfficientNetB4 (0.85)	EfficientNetB0 (0.85)
5th	DenseNet169 (1.05)	ResNet152 (0.85)	EfficientNetB5 (0.85)	ResNet50 (0.85)

Three-model based weighted averaging method used rank 1 to 3 models while five-model based weighted averaging method used rank 1 to 5 models, with their corresponding weightages in the parenthesis.

reduction observed when all base learners were coming from the same CNN architecture. For instance, when EfficientNetB0 to B4 with similar preferences were selected for 5× magnification factor case, the majority voting approach still managed to yield the second highest Acc, followed by the weighted averaging approach. However, the majority voting approach did not perform as good as the other ensemble approaches when the magnification factors increased.

For unweighted averaging ensemble method, the performance was stable along the changes of magnification factors, and it was highly dependent on the number of base CNN models involved. Ensemble-UA3 model acquired the second highest Acc among the three-CNNs ensemble techniques in all magnification factors. But when the number of base CNN models increased to five, its performance dropped drastically for any of the scenario, and even yielded a lower Acc than the Ensemble-UA3 model. This is because the first-best and the fifth-best selected CNN models tend to have a great difference in performance. With that, the favorable performance from the first-best CNN model would be compromised, resulting in a poorer result in overall.

Meanwhile, weighted averaging approach can overcome the limitations for both majority voting and unweighted averaging approaches. Unlike majority voting, weighted averaging was more robust, and it could maintain its good performances as the magnification factor or number of images increased. Apart from the 1.25× magnification factor setting, in which the Ensemble-WA5 acquired the second-best results, the Ensemble-WA5 model outperformed the other ensembled models in the other three magnification factor settings. The customizable weights enable the weighted averaging approach to adjust the weight coefficients for a particular CNN model, depending to its performance, until the best result can be obtained. The weightage used by the weighted averaging ensemble method for each magnification factor is tabulated in *Table 6*. The only drawback of weighted averaging ensemble approach is it requires supervision and manual adjustment on the weights for better results. Considering its reliability as a diagnosis system, three-CNN based weighted averaging ensembled model in 1.25× magnification factor was the only ensembled model which slightly failed the test as a reliable diagnosis system, with a 0.03% Sp scored below the requirements.

Stacking ensemble technique is the only non-statistical based ensemble approach tested in this study, and it resulted a moderate performance among the selected ensemble approaches. Although its improvement in Acc when increased the base CNN models from three to five was not always the greatest, its improvement was consistent with the changes of magnification factors. Still, it did not outperform either majority voting or weighted averaging methods, probably because it involved two levels of learning process, and the second learning process with validation set was only 20% of the entire dataset, making the meta-learner did not have sufficient training in performing ensemble learning with the selected CNN models. K-fold cross validation approach may be more favorable for stacking technique than the hold-out validation approach used in this study as it can train the meta-classifier with more different images.

Comparative analysis of ensembled models with the best CNN models

The comparison of the proposed ensembled models with individual CNN models in various magnification factors shows that the ensembled models mostly outperformed the

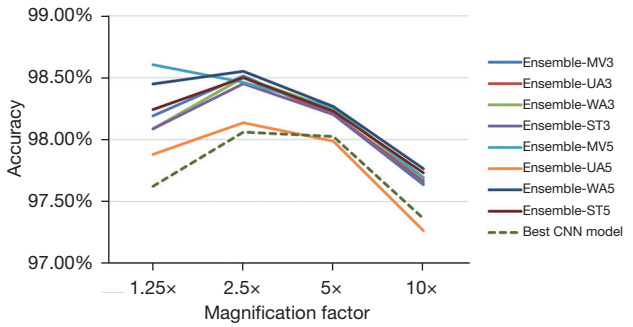

Figure 6 Comparison in accuracy between ensembled models and the best CNN model for each tested magnification factor. CNN, convolutional neural network; MV, majority voting; UA, unweighted averaging; WA, weighted averaging; ST, stacking.

individual CNN models. Except for Ensemble-UA5 model, all ensemble methods acquired a better Acc than the best performing individual CNN model in every magnification factors as shown in *Figure 6*. Therefore, the Ensemble-UA5 model would not be included for the following discussions. By comparing with the best individual CNN model, the Ensemble-WA5 model shows the best Acc improvement in overall magnification factor settings by at most 0.83%. On average, ensembled models with three-CNN models resulted in a 0.35% improvement in Acc, while models with five-CNN models resulted in a 0.46% Acc improvement.

However, ensembled models did not always yield better results than the best individual CNN model in all metrics. Instead, a lower Sp and Pr were possible to happen as a trade-off for the high Re in the ensembled models. For instance, even though the Ensemble-WA3 model at 1.25× magnification factor resulted a high improvement of 1.67% in Re, it had a drop of 2.11% and 0.93% in Sp and Pr, respectively. In other words, the ensembled model would rather have more FPs than having the occurrence of FNs. As the magnification factors increases, the relationship between these three evaluation metrics (Re, Sp, and Pr) is more stable.

For an imbalanced dataset, the F1 of an automated diagnosis system is considerably important to emphasis the harmonic mean of Pr and Re without any biases. The investigated ensembled models show a 0.13% to 0.73% improvement in F1 as compared to the maximum F1 reported by the individual CNN models. In other words, Pr and Re are balanced to each other instead of improved at the expense of the other. Likewise, the ensembled models significantly improved another measure for imbalanced dataset, MCC, as compared to the CNN models. The improvement of 0.40% to 2.23% in MCC shows that the ensembled models are able to predict for better results in all the entries of the confusion matrix categories.

Performance comparison with the state-of-the-art results

Table 7 compares the best proposed ensembled model from different magnification factor settings with the techniques reported in the literature. The results suggest that the proposed Ensemble-WA5 model shows significant improvement in almost all evaluation metrics compared to the existing literature. Comparing to the work by Azuaje *et al.* (44), the Acc, Sp, F1, and AUC of the best proposed ensembled model are higher by 4%, 5%, 7%, and 6% for the best case, respectively. The Re of our ensembled model, on the other hand, is slightly inferior by 1% to 2%. Meanwhile, the best proposed ensembled model outperformed the EfficientNetB0 from Koo *et al.* (45) in every evaluated metric, especially in Sp, where the improvement is 4% when compared with the same magnification factor of 5×. Overall, our suggested technique outperformed the state-of-the-art results and had superior classification results in four out of the five evaluated metrics.

Discussion

The existence of renal cancer has a detrimental impact on patients' quality of life. Additionally, renal cancer patients diagnosed at late stages also have to endure a substantial economic burden for a very low survival rate. Early detection of renal cancer could aid in the development of an appropriate treatment plan, which leads to a higher survival rate. The emergence of deep machine learning can assist and facilitate the pathologists in effectively analyzing and classifying the cancer biopsies. In this work, we have demonstrated multiple deep machine learning models and ensemble learning techniques can discriminate the renal tissues into normal and tumor classes effectively and efficiently.

The proposed pipeline for renal cancer classification first applied segmentation and image patching with different downsampling factors to generate histopathological images with a smaller dimension. Five different pre-trained CNN architectures with ImageNet pre-trained weights are fine-tuned, and the diversity between architectures has been demonstrated with Grad-CAM. EfficientNet has resulted the best overall performance in all the evaluated metrics out

Table 7 Model comparison on the CPTAC-CCRCC dataset with the best result proposed ensemble model in this study and the literature

Model	Magnification factor	Acc	Sp	Re	F1	AUC
VGG16 (44)	–	0.95	0.93	**1.00**	0.92	0.92
EfficientNetB0 (45)	5×	0.97	0.94	0.98	0.98	0.96
Proposed Ensemble-WA5	1.25×	0.98	0.97	0.99	**0.99**	**0.98**
	2.5×	**0.99**	**0.98**	0.99	**0.99**	**0.98**
	5×	0.98	**0.98**	0.99	**0.99**	**0.98**
	10×	0.98	0.96	0.98	0.98	0.97

Since the magnification factor of the literature was not reported, four magnification factors from this study are compared. The results are reported in two decimal places for fair comparison and best results of the metrics are bolded. CPTAC-CCRCC, Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma; WA, weighted averaging; Acc, accuracy; Sp, specificity; Re, recall; F1, F1-score; AUC, area under the curve.

of the five architectures. Although EfficientNet managed to obtain a relatively high score in most of the evaluated metrics, the Sp score is consistently the lowest. This might be owing to the dataset's imbalanced distribution, since the tumor class consists of roughly 67% of the dataset, resulting in an overfitting classification to the majority class. In other words, because of the imbalanced nature of the dataset, every CNN model, including the best performing model from EfficientNet, may tend to predict a non-tumor renal image as tumorous.

Instead of just having a single CNN model for classifying renal cancer as in the literature, we have employed four different ensembled models approaches with different number of base learners. Weighted averaging with five-CNN model (Ensemble-WA5) is the best performing ensemble approach in general. It shows more improvement towards the best performing individual CNN models in most of the evaluated metrics, and it is more robust towards the changes in magnification factors. Moreover, the improvement in Acc, Sp, F1, and AUC compared to the state-of-the-art results suggests that our best-proposed ensemble technique performs better on the same renal histopathological dataset. Although we have inferior Re than the benchmark, the best proposed ensembled model managed to substantially increase the score in other four compared metrics and meet the requirement as a reliable diagnosis system with just 1% sacrificial in Re.

The feasibility and efficiency of utilizing deep machine learning and ensemble learning to classify the non-annotated renal histopathological images have been demonstrated. The use of slide-level label is suitable for classification of histopathological images in various magnification factors. Moreover, the developed models are robust against the changes in magnification factors. With this implemented in the actual clinical site, the pathologists can get the clinical diagnoses for different magnification factors effectively without performing annotation on the histopathological images; thus, they can spend more time on treatment planning and making better decisions. However, the designed pipeline in this study still requires minimal supervision and manual adjustments, such as removing undesirable generated image patches and adjusting weight coefficients for weighted averaging ensemble model. Furthermore, the computational costs are higher due to multiple CNN models are required to perform the ensemble learning to achieve the reported state-of-the-art performance. In the future, we intend to develop a fully automated histopathological image classification pipeline which is more feasible and suitable for real-life clinical practice. This can be done by having artifacts detection algorithms for masking or auto-removal of the artefacts before converting the WSI into patches and implement feedforward neural network to learn to assign the most suitable weight coefficient to each model's output based on how well that model's output matches the true label. Parallel computing will be considered in the future instead of sequential computing approach as has been done in this work, to accelerate the training, validating, and testing process. This is especially important when more base models are required to perform the proposed ensemble learning model. With the success in this study, we will also test the proposed work in other renal histopathological image datasets in the future to assess the robustness, Acc, and reliability of our proposed method.

Conclusions

In the present study, we have developed deep CNN based heterogeneous ensemble models to classify the renal histopathology images into normal and tumor classes effectively. In this regard, multiple CNN models from five selected architectures were fine-tuned and trained with several magnification factors. The selected architectures showed sufficient diversity and Acc in discriminating tumor tissues. EfficientNet was found to be the best architecture as most of its variants were able to achieve an outstanding classification performance. The classification performance of the best-proposed ensembled model: Ensemble-WA5, had the state-of-the-art results in Acc (99%), Sp (98%), F1 (99%) and AUC (98%) but slightly inferior Re (99%) when compared to the published literature. It fulfilled the requirement of a reliable diagnosis system and the results showed that it was feasible to use deep machine learning algorithms to classify non-ROI-annotated renal histopathological images with different magnification factors to assist pathologists in manual inspection and make pathology diagnosis more effective and efficient. In the future, this work can be extended to automate the histopathology analysis workflow in clinical sites to assist pathologists in the manual inspection and making the diagnosis process more efficient, and less misdetection, misdiagnosis and inter-pathologist variation.

Acknowledgments

Some parts of the content in this article are from the first author's thesis.
Funding: This work was supported by Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (No. FRGS/1/2021/ICT02/UTAR/02/2); Universiti Tunku Abdul Rahman (UTAR) under the UTAR Strategy Research Fund (No. IPSR/RMC/UTARSF-SR/PROJECT 2021-C1/001); and UTAR Research Fund (No. IPSR/RMC/UTARRF/2020-C1/S07).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-46/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-46/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
2. Padala SA, Kallam A. Clear Cell Renal Carcinoma. 2023.
3. Motzer RJ, Jonasch E, Michaelson MD, Nandagopal L, Gore JL, George S, et al. NCCN Guidelines Insights: Kidney Cancer, Version 2.2020. J Natl Compr Canc Netw 2019;17:1278-85.
4. Moch H. The WHO/ISUP grading system for renal carcinoma. Pathologe 2016;37:355-60.
5. Boumaraf S, Liu X, Zheng Z, Ma X, Ferkous C. A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images. Biomed Signal Process Control 2021;63:102192.
6. Samaratunga H, Gianduzzo T, Delahunt B. The ISUP system of staging, grading and classification of renal cell neoplasia. J Kidney Cancer VHL 2014;1:26-39.
7. Li X, Li C, Rahaman MM, Sun H, Li X, Wu J, Yao Y, Grzegorzek M. A Comprehensive Review of Computer-aided Whole-slide Image Analysis: from Datasets to Feature Extraction, Segmentation, Classification and Detection Approaches. Artif Intell Rev 2022;55:4809-78.
8. Yari Y, Nguyen H. A State-of-the-art Deep Transfer

Learning-Based Model for Accurate Breast Cancer Recognition in Histology Images. In: 2020 - BIBE 2020 - The 20th IEEE International Conference on Bioinformatics and Bioengineering. Cincinnati, OH, USA; 2020:900-5.
9. Hameed Z, Zahia S, Garcia-Zapirain B, Javier Aguirre J, María Vanegas A. Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models. Sensors (Basel) 2020;20:4373.
10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
11. Xie J, Liu R, Luttrell J 4th, Zhang C. Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Front Genet 2019;10:80.
12. Tham ML, Wong YJ, Kwan BH, Owada Y, Sein MM, Chang YC. Joint Disaster Classification and Victim Detection using Multi-Task Learning. In: 12th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, UEMCON 2021. New York, NY, USA; 2021:407-12.
13. Wong YJ, Tham ML, Kwan BH, Gnanamuthu EMA, Owada Y. An Optimized Multi-Task Learning Model for Disaster Classification and Victim Detection in Federated Learning Environments. IEEE Access 2022;10:115930.
14. Sarwindaa D, Paradisa RH, Bustamam A, Anggia P. Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. Procedia Comput Sci 2021;179:423-31.
15. Kallipolitis A, Revelos K, Maglogiannis I. Ensembling Efficient Nets for the Classification and Interpretation of Histopathology Images. Algorithms 2021;14:278.
16. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. arXiv:1409.4842v1 [Preprint]. 2014 (cited 2022 Jan 7). Available online: https://arxiv.org/abs/1409.4842v1
17. Datta Gupta K, Sharma DK, Ahmed S, Gupta H, Gupta D, Hsu CH. A Novel Lightweight Deep Learning-Based Histopathological Image Classification Model for IoMT. Neural Process Lett 2023;55:205-28.
18. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861v1 [Preprint]. 2017 (cited 2021 Dec 27). Available online: https://arxiv.org/abs/1704.04861v1
19. Ikromjanov K, Bhattacharjee S, Hwang YB, Kim HC, Choi HK. Multi-class Classification of Histopathology Images using Fine-Tuning Techniques of Transfer Learning. J Korea Multimed Soc 2021;24:849-59.
20. Voon W, Hum YC, Tee YK, Yap WS, Salim MIM, Tan TS, Mokayed H, Lai KW. Performance analysis of seven Convolutional Neural Networks (CNNs) with transfer learning for Invasive Ductal Carcinoma (IDC) grading in breast histopathological images. Sci Rep 2022;12:19200.
21. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep Learning for Identifying Metastatic Breast Cancer. arXiv:1606.05718v1 [Preprint]. 2016 (cited 2021 Dec 27). Available online: https://arxiv.org/abs/1606.05718v1
22. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
23. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging 2016;35:1299-312.
24. Kandel I, Castelli M. How Deeply to Fine-Tune a Convolutional Neural Network: A Case Study Using a Histopathology Dataset. Appl Sci 2020;10:3359.
25. Yong MP, Hum YC, Lai KW, Lee YL, Goh CH, Yap WS, Tee YK. Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning. Diagnostics (Basel) 2023;13:1793.
26. Mittal S. Ensemble of transfer learnt classifiers for recognition of cardiovascular tissues from histological images. Phys Eng Sci Med 2021;44:655-65.
27. Paladini E, Vantaggiato E, Bougourzi F, Distante C, Hadid A, Taleb-Ahmed A. Two Ensemble-CNN Approaches for Colorectal Cancer Tissue Type Classification. J Imaging 2021;7:51.
28. Chen CL, Chen CC, Yu WH, Chen SH, Chang YC, Hsu TI, Hsiao M, Yeh CY, Chen CY. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nat Commun 2021;12:1193.
29. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J 2018;16:34-42.
30. Rathore S, Niazi T, Iftikhar MA, Chaddad A. Glioma Grading via Analysis of Digital Pathology Images Using Machine Learning. Cancers (Basel) 2020;12:578.
31. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep

neural networks. Sci Rep 2019;9:3358.
32. Celik Y, Talo M, Yildirim O, Karabatak M, Acharya UR. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. Pattern Recognit Lett 2020;133:232-9.
33. Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, Coiera E, Di Ieva A. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020;10:7733.
34. Im S, Hyeon J, Rha E, Lee J, Choi HJ, Jung Y, Kim TJ. Classification of Diffuse Glioma Subtype from Clinical-Grade Pathological Images Using Deep Transfer Learning. Sensors (Basel) 2021;21:3500.
35. Phan NN, Hsu CY, Huang CC, Tseng LM, Chuang EY. Prediction of Breast Cancer Recurrence Using a Deep Convolutional Neural Network Without Region-of-Interest Labeling. Front Oncol 2021;11:734015.
36. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5:555-70.
37. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301-9.
38. The Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma Collection (CPTAC-CCRCC). (cited 2021 Oct 7). Available online: https://doi.org/10.7937/K9/TCIA.2018.OBLAMN27
39. Muñoz-Aguirre M, Ntasis VF, Rojas S, Guigó R. PyHIST: A Histological Image Segmentation Tool. PLoS Comput Biol 2020;16:e1008349.
40. Luo L, Xiong Y, Liu Y, Sun X. Adaptive Gradient Methods with Dynamic Bound of Learning Rate. arXiv:1902.09843v1 [Preprint]. 2019 (cited 2021 Dec 6). Available online: https://arxiv.org/abs/1902.09843v1
41. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. arXiv:2104.02395v3 [Preprint]. 2022 (cited 2022 Sep 1). Available online: https://arxiv.org/abs/2104.02395v3
42. Zhou ZH. Machine Learning. Singapore. Springer Nature Singapore Pte Ltd., 2021.
43. Kandel I, Castelli M, Popovi A. Comparing Stacking Ensemble Techniques to Improve Musculoskeletal Fracture Image Classification. J Imaging 2021;7:100.
44. Azuaje F, Kim SY, Perez Hernandez D, Dittmar G. Connecting Histopathology Imaging and Proteomics in Kidney Cancer through Machine Learning. J Clin Med 2019;8:1535.
45. Koo JC, Hum YC, Lai KW, Yap WS, Manickam S, Tee YK. Deep Machine Learning Histopathological Image Analysis for Renal Cancer Detection. In: 8th International Conference on Computing and Artificial Intelligence (ICCAI). Tianjin, China; 2022:657-63.

Cite this article as: Koo JC, Ke Q, Hum YC, Goh CH, Lai KW, Yap WS, Tee YK. Non-annotated renal histopathological image analysis with deep ensemble learning. Quant Imaging Med Surg 2023;13(9):5902-5920. doi: 10.21037/qims-23-46

