Peer Review File

Article information: https://dx.doi.org/10.21037/tp-23-508

Response To Reviewer A

Comment 1: I don't see any annotations on how many images were collected for one tumor - it should be clearly defined (1 image / many images + their statistics)
Reply 1: Thanks to the reviewer for your first suggestion. In the section "Image acquisition and segmentation", "ROIs were constructed by manually drawing along the tumor edge on the largest cross-sectional plane" is described. Actually, we drew 1 two-dimensional ROI image on the largest cross-section of CMP and NP.
Changes in the text: ROIs were constructed by manually drawing along the tumor edge on the largest cross-sectional plane and one image was collected from the CMP or NP respectively.
Thanks again for your suggestion.

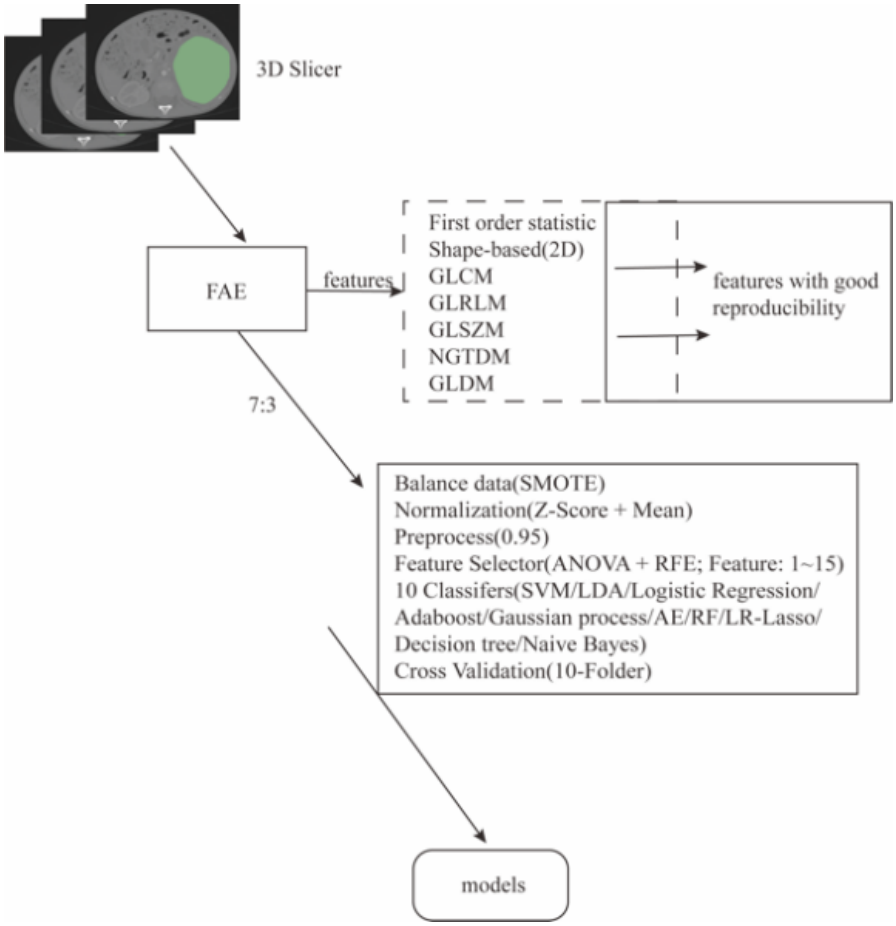
Comment 2: statistics of tumor sizes are missing (were the WT and non-WT databases significantly different?), they should also be added to the appropriate tables
Reply 2: We sincerely appreciate your second opinion. We have added tumor size to our table and do statistical analysis. The basis for selecting the cut-off point for maximum diameter was determined based on the median maximum diameter of all tumors.
Thanks again for your opinion.
Changes in the text: Add the relevant data to the table 1 and table 2.

Comment 3: the article requires thorough linguistic corrections (sentence structure), some sentences are excessively long + numerous repetitions (especially in the introduction), without providing additional information. Additionally numerous linguistic errors.
Reply 3: Thanks for your recommendation. We accept the editor's advice, choosing "AME Editing Service".
Thanks again for your recommendation.
Changes in the text: None

Comment 4: in my opinion, there is a lack of an example image summarizing the structure of the network, along with an exemplary outline of the tumor and examples / all features according to which the network was trained.
Reply 4: Thank you for your opinion. We drew an image (figure 1), along with an

exemplary outline of the tumor and all features and all classifiers. Thanks again.
Changes in the text:

Figure 1



Comment 5: table 1 does not add much to the article (the time of contrast administration / the patient's anatomical conditions may be even more important for the image vs. the voltages itself) - it proposes a collective table of all cases with precise statistics (age, gender, tumor size, etc.)
Comment 6: table 2 lists all voltages again for no reason - requires rebuilding. it can be combined with table no. 1.
Comment 7: table 3 - as above
Comment 8: in my opinion, tables 4 and 5 should be combined
Reply 5~8: We sincerely thank you for allowing us to modify tables. We have combined tables 1, 2, and 3 into one table and have added data on tumor size. And, we also have combined tables 4 and 5.
Thank you very much!
Changes in the text: tables 1 and 3.

Comment 9: the discussion does not mention the study results (Zhu Y, Li H, Huang Y, Fu W, Wang S, Sun N, Dong D, Tian J, Peng Y. CT-based identification of pediatric non-Wilms tumors using convolutional neural networks at a single center. Pediatr Res. 2023 Sep;94(3):1104-1110) examining the same issue - should at least refer to the results and propose methods to improve its effectiveness in the future

Reply 9: We are grateful for the suggestion. According to your suggestion, we have read this article in detail again, and then made the corresponding summary in the discussion section.

Thanks again.

Changes in the text: In terms of building a model to distinguish Wilms tumors from non-Wilms tumors, Zhu et al(3) built a deep learning (DL) model (ResNet34) to achieve this. In contrast to their study, we carefully drew the ROI along the edge of the tumor to avoid the influence of peritumoral renal parenchyma and fat. However, they choose 1.2ROI to extract features. Fine manual drawing of ROI along the tumor margin may be beneficial to the final result of the models. Whether it is ML or DL, the performance of models is most closely related to the sample size. To improve the performance of our model, the sample size should be further increased in the future.

Comment 10: the discussion is not well written and should be rebuilt, and should also mention today's standards - it does not mention the possibility of biopsy

Reply 10: Thank you very much for your key advice. For this suggestion, we first carefully searched and read the relevant literature, made the corresponding summary, and then added it to our discussion section, and made the corresponding modifications and adjustments. Thank you very much for your suggestion.

Changes in the text:

If the type of renal tumor in children can be diagnosed by imaging before surgery, children with renal tumors will benefit. There are differences in the treatment of pediatric renal tumors in different regions. For example, in North America (The Children’s Oncology Group, COG), it is advocated that tumor resection should be performed first, and subsequent treatment should be carried out after the pathology is confirmed. In Europe (The International Society of Paediatric Oncology, SIOP), preoperative chemotherapy to reduce tumor staging and the risk of rupture is followed by operation. For children between the ages of 6 months and 9 years, the standard chemotherapy regimen for Wilms tumor is generally adopted, and when non-Wilms tumor is suspected, the chemotherapy regimen is determined based on the results of biopsy. For children older than 10 years with an uncertain clinical diagnosis, the biopsy is generally performed first(11-13). Wilms tumors can be distinguished from non-Wilms tumors by non-invasive methods (imaging), which is beneficial both for the plan of direct resection of the tumor and the plan of further treatment after biopsy, and biopsy

may cause needle tract metastasis and increase the clinical stage.

Comment 11: in "Key findings" the sentence "ML models are good at identifying non-Wilms tumors from Wilms tumors in children with excellent results, which is better than the human." is too strong, there is no reference to it

Reply 11: We are grateful for the suggestion. We have changed this sentence to “ML models are good at identifying non-Wilms tumors from Wilms tumors in children with relatively good results, which is may better than the human”.In our study, the performance of the model was better than that of preoperative diagnosis, so we propose this idea. Since this is a single-center study, this idea needs to be further verifiedThanks again.

Changes in the text: ML models are good at identifying non-Wilms tumors from Wilms tumors in children with relatively good results, which is may better than the human.

Comment 12: citations require checking (e.g. for reference 21, there is no information in the content of the article confirming the given age data)

Reply 12: We would like to thank the reviewers for your last comment. We have checked all citations and inappropriate citations have been replaced.

Thank you for your advice and guidance, which is extremely valuable.

Changes in the text: the age distribution of each renal tumor in children is different(24,25)

24. Ahmed HU, Arya M, Levitt G, et al. Part I: Primary malignant non-Wilms' renal tumours in children. Lancet Oncol 2007;8:730-7.

25. Wilms tumour. Nat Rev Dis Primers 2021;7:76.

Response To Reviewer B

Overall comment: 1. The author should provide sample CT images of the renal tumors studied in this paper. 2. Provide more information about CMP and NP CT protocol. What were the enhancing techniques used in acquiring these images?

Reply : I am very grateful for your comment on the manuscript. 1. The sample images of this study will be provided. And the third and fourth rule of “Data Sharing Statement”, which is attached, has been amended. 2. Detailed information on CT has also been provided.

Thank you for your advice and guidance, which is extremely valuable.

Changes in the text: 1. “Data Sharing Statement”:

What data in particular will be shared?

CT images (ROI). Futures which extracted by the software.

Any other documents will be share? Such as study protocol, statistical analysis plan, informed consent form, clinical study report, analytic code.

Study protocol, statistical analysis plan and software will also be shared if requested.

2. A plain scan was performed before the enhanced CT scan. The contrast agent was ioversol (350mg/mL), and the injection dose was 1.5mL/kg. The CMP was scanned for about 15~30s after the injection of the contrast agent, and the NP was scanned for about 60~90s.

Comment 1: Line 240: In our research, 82 patients, 77 images of CMP, and 81 images of NP were included to make an analysis, and Table 1 confirms this. Why are 5 patients in CMP and 1 patient in NP excluded? The author is asked to provide significant differences between CMP and NP test datasets. It is essential because you are comparing CMP and NP results, but your test dataset appears different.

Reply 1: Thanks for your suggestion. In our study, 5 patients in CMP and 1 patient in NP were excluded. The reason was that CMP was missing in 5 patients and NP was missing in 1 patient. Accordingly, a table was drawn to verify whether different data sets would have an impact on the CMP and the NP. These data are also summarized in Table 1 and Table 2.

Thank you very much!

Changes in the text:

Table 1

	Non-Wilms tumor	Wilms tumor	Value	P
CMP	19	58	0.032	0.858
NP	19	62		

Table 2

Variables	CMP	NP	Value (χ^2)	P

Age (months)			1.611	0.204
≥26	35	45		
<26	42	36		
Gender			0.002	0.964
Female	33	35		
Male	44	46		
Side			0.057	0.812
Left	48	49		
Right	29	32		
Maximum diameter(cm)			0.154	0.695
≥11	26	26		
<11	51	55		

Comment 2: Line 248-250: CMP model results are compared with preoperative image diagnosis. Please both CMP models' accuracy for comparison purposes.
Reply 2: Thanks for your recommendation. We have compared the accuracy of all CMP models with preoperative image diagnosis.
Thanks again for your recommendation.
Changes in the text: The accuracy of CMP models ranged from 0.630 to 0.833, and our all CMP models were better than the accuracy (0.592) of preoperative image diagnosis.

Comment 3: Line 256: It says about 2 but the first option is not mentioned.
Reply 3: We sincerely appreciate your suggestion. The first option has been added in place.
Once again, thank you very much.
Changes in the text:
1. There are many kinds of non-Wilms' tumors, including CCSK, renal, RCC, MRTK, CMN, multilocular cystic nephroma, and other types.

Comment 4: Line 257: Please provide some visualization proof to support this claim. Sample images will work to show that CT value of the renal parenchyma is higher than that of the tumor.
Reply 4: We are grateful for the suggestion. We measured the mean CT values of renal parenchyma and renal tumor in the maximum cross-sectional section of the tumor in the CMP, and then calculated the mean values of renal parenchyma and renal tumor in 77 images. The mean CT value was 120HU for renal parenchyma and 45HU for renal

tumor.
Thanks again.
Changes in the text: Renal tumors in children are large and often accompanied by internal necrosis and cystic degeneration. In CMP, the CT value of the renal parenchyma is higher than that of the tumor (mean CT value: 120HU VS. 45HU), thus potentially masking the tumor.

Comment 5: Line 265-266: CMP is better than NP in discrimination, but NP is less affected by various factors and more stable. Please elaborate on what type of discrimination, and what type of factors you are considering when concluding.
Reply 5: We would like to thank the reviewers for your last comment. The constructed CMP model is superior to the NP model in the differential diagnosis of Wilms tumor and non-Wilms tumor. The NP model may be less affected by perirenal fat and renal parenchyma and thus more stable. This is mentioned in the discussion section of our manuscript.
Special thanks to you for your good comments.
Changes in the text: None

Response To Reviewer C

Comment 1: There is need for extensive editing of the manuscript, hence can not be published in the current form.
Reply 1: I am very grateful for your comment on the manuscript. We accept the editor's advice, choosing “AME Editing Service”.
Thank you for your advice and guidance, which is extremely valuable.
Changes in the text: None

Comment 2: In the abstract line 50 and 51, the CMP results are different from the results and discussion. Why is this so?
Reply 2: We sincerely appreciate your suggestions. First, in the discussion section, the value of SPE should be consistent with the abstract section. Secondly, for the SEN values of CMP (0.539 vs 0.923) in the abstract, results, and discussion sections, the pipelines producing these two values are different. The discussion section is the highest value of all pipelines (Mean-PCC-RFE-1-LR), while the abstract section corresponds to the SEN value of pipeline (Mean-PCC-ANOVA-5-AE) corresponding to the highest AUC value.
Thanks again.
Changes in the text: In the CMP model, we received the highest performance of AUC

(0.792), ACC (0.833), SEN (0.923), and SPE (0.927), where the highest AUC, ACC and SPE were generated by the Mean-PCC-ANOVA-5-AE pipeline model and the highest SEN was generated by the Mean-PCC-RFE-1-LR pipeline model.

Comment 3: In the image segmentation section, the authors indicated that they are delineating the ROI of the largest tumor slice however they saved it in NIFTI format, line 118.

Reply 3: We are grateful for the suggestion. We used 3D slicer software to draw the ROI on the largest cross section of the tumor on the CT image, and then saved the drawn ROI in NIFTI format for the purpose that the FAE software could do the next step. FAE software cannot directly process DICOM format files.

Thanks again.

Changes in the text:

Firstly, the DICOM file was obtained and input into the open-source 3D Slicer software (version 5.2.2)(8) based on the "Segment Editor" module to draw the two-dimensional ROI on CMP and NP, and the drawn ROI was output in NIfTI format for the next operation.

Comment 4: In line 120, you indicated that the SPSS software was used to delineate the ROI, however, as far as I know SPSS is a statistical tool and not a delineation tool. Elaborate.

Reply 4: We sincerely appreciate your suggestion. You misunderstood the meaning due to our poor explanation. The random sampling function of SPSS software was used to randomly extract 10% of all the images (CMP 8/77; NP:8/81), and then 3D slicer software was used to redraw the ROI of the 10% of the images to screen the highly reproducible features.

Once again, thank you very much.

Changes in the text: Author HHS drew all ROIs independently, meanwhile, author XQW used SPSS software to select 10% of images randomly and then to delineate ROI using 3D slicer software.

Comment 5: In the research you indicated you used 10 classifiers, however, in the results and even the discussion the performance of only a single classifier has been given and even for that one classifier we have not been given the name. Justify.

Reply 5: We would like to thank the reviewers for your last comment. The 10 classifiers we selected are SVM, LDA, LR, Adaboost, Gaussian process, AE, RF, LR-Lasso, Decision tree, and Naive Bayes. A total of 600 pipelines were formed according to different normalization methods (Mean/Z-score), dimensionality reduction methods (PCC), feature selection methods (ANOVA/RFE to select 15 features), and 10 different

classifiers. The final expression of our model is presented in the form of pipeline, such as “Mean-PCC-ANOVA-5-AE”, where “AE” is one of the 10 classifiers. In addition, Figure 2 and Figure 4, for example, represent the top five pipelines according to the AUC metric. In addition, the main purpose of this paper was to build machine learning models to identify the possibility of Wilms tumor and non-Wilms tumor, so only the best performing among 600 pipelines were sought and the performance of the 10 classifiers was not compared. In the future, our research direction may extend to the influence of different classifiers on the model.

Special thanks to you for your good comments.

Changes in the text:

Mean-PCC-ANOVA-5-AE model →Mean-PCC-ANOVA-5-AE pipeline model

Mean-PCC-ANOVA-2-LR model →Mean-PCC-ANOVA-2-LR pipeline model