Peer Review File

Article Information: https://dx.doi.org/10.21037/jgo-23-587

Reviewer A

First of all, my major concern regarding the methodology of this study is that the patches used for the development and validation of the discrimination model is from 10 patients only, potentially limiting the external validity of the model.

Second, the abstract needs some revisions. The introduction did not explain why the U-Net Deep Learning Segmentation Model is potentially accurate and what the limitations of available algorithms are. The methods need to describe the patient samples in the development, internal validation, and external validations. The authors need to describe the statistical indicators for assessing the accuracy of the model. The results need to briefly summarize the clinical characteristics of the patient samples and important accuracy indictors in the external validation samples should be reported such as AUC, sensitivity and specificity. The conclusion needs more detailed comments for the clinical implications of the findings.

Third, in the introduction of the main text, the authors need to review the accuracy and limitations of other algorithms such as VGG16, VGG19, Resnet50, Xception and scratch-based models, and analyze why U-Net Deep Learning Segmentation Model is potentially accurate.

Fourth, in the methodology of the main text, please clearly and accurately describe the clinical research design of this study and explain the appropriateness of the small clinical sample for the development of the model, as well as the clinical samples in the two external validation datasets. In the statistics, please provide the threshold values of these accuracy indicators for a good discrimination model.

Finally, please consider to review and cite some relevant papers, which may enrich the content of this paper: 1. Tokai Y, Yoshio T, Fujisaki J. Development of artificial intelligence for the detection and staging of esophageal cancer. Ann Esophagus 2023;6:3. 2. Cao F, Chen G, Su W, Zhang Z, Fu Q, Zhou D, Dai Z. Endoscopic ultrasound-guided fine needle aspiration for smooth benign appearing malignant esophageal stricture: a cross-sectional study. J Thorac Dis 2022;14(6):2112-2121. doi: 10.21037/jtd-22-584. 3. Rezaeijo SM, Jafarpoor Nesheli S, Fatan Serj M, Tahmasebi Birgani MJ. Segmentation of the prostate, its zones, anterior fibromuscular stroma, and urethra on the MRIs and multimodality image fusion using U-Net model. Quant Imaging Med Surg 2022;12(10):4786-4804. doi: 10.21037/qims-22-115.

Reply: Thank you for your critical comment. First, we also acknowledged the inherent limitation of our study, namely the small sample size, which may restrict the applicability of the findings, especially in the context of a diverse condition like tumors. This limitation has been discussed in our manuscript, and the efficacy of the models in the external validation is significantly lower. Nonetheless, we contend that our study presents a viable foray into deep learning image segmentation on esophageal pathological images. Moving forward, if deemed necessary, our future endeavors will involve undertaking more comprehensive image segmentation on a larger dataset to further enhance the precision of our models.

Second, revisions have been made on the abstract. However, due to the limitation of content, we have made corresponding answers to some suggestions in the text rather than in the abstract. One important limitation of available algorithms is that large biomedical datasets containing thousands of training images are still required for the complex outputs. And the U-Net network showed precise segmentations with very few training images for biomedical images, and therefore it is assumed that U-Net Deep Learning Segmentation Model is potentially accurate. The cohort and statistical indicators were supplemented in the methods. Clinical characteristics as well as important indictors in the external validation were summarized in the result. And conclusion has been modified.

Third, various networks structures including VGG16、VGG19、Resnet50、Xception has been proposed for medical images and achieve precise segmentation. However, the development of traditional CNN network requires a large image cohort to ensure efficacy, while the U-Net network could segment biomedical images with very few training images (page 6, line 97-101). The reasons for the accuracy of U-Net Deep Learning Segmentation Model might be the inherent of the network.

Forth, the clinical samples in the two external validation datasets were supplemented in the table 2. Previous studies have not defined the threshold values of accuracy indicators for a good discrimination model, in the discussion, we assumed a value of 80% as good discrimination (page 13, line 255-257). And therefore, the models developed in this study exhibited satisfactory results. Finally, thanks again for reviewer's kind advice, and we have cited some relevant papers to enrich the content of this paper.

Reviewer B:

The paper titled "The Development and Validation of Pathological Sections based U-Net Deep Learning Segmentation Model for the Detection of Esophageal Mucosa and Squamous Cell Neoplasm" is interesting. The models developed in this study exhibited satisfactory results, paving the way for their potential deployment on standard computers and integration with other artificial intelligence models in clinical practice in the future. However, there are several minor issues that if addressed would significantly improve the manuscript.

1) What are the biggest strengths and weaknesses of this research model? What is the biggest problem faced? Suggest adding relevant content.

2) The number of patients and WSIs included was relatively small. It is recommended to increase the sample size to avoid affecting the generalizability of the model.

3) The introduction part of this paper is not comprehensive enough, and the similar papers have not been cited, such as "Progress on deep learning in digital pathology of breast cancer: a narrative review, J Gland Surg, PMID: 35531111". It is recommended to quote the article.

4) How to compare the recognition accuracy of different network structures? It is recommended to add relevant content.

5) In addition to the method of this study, what other methods can be used to achieve this effect? Please analyze based on the literature.

6) What guidance can this study provide for the early diagnosis and adjuvant treatment in esophageal squamous cell carcinoma? It is suggested to add relevant contents.

Reply:

Thanks for your valuable commentary.

1) The biggest strength of our study is the development of ESCC segmentation models from small cohort, and models showed acceptable effectiveness in both internal and external validation (page 14, line 274-279). The weaknesses and biggest problem faced is that the number of patients and WSIs included in the

study was relatively small, which might affect the generalizability of the model (page 14, line 290-293).

2) We fully agree with reviewer's advice. Although the AUC of our models might be acceptable, the number of patients and WSIs could limit generalizability. In further study, we plan to increase the sample size the achieve greater generalizability.

3) Thanks again for reviewer's helpful suggestions, and we have cited the relevant paper to enrich the comprehensiveness of our manuscript.

4) We also tried to find a parameter to compare the recognition accuracy of different network structures. However, to the best of knowledge, there are no acknowledged comprehensive parameters for the model comparison, with IOU most acceptable in the evaluation of recognition accuracy. Some previous articles also used comprehensive curves of speed performance and accuracy in the assessment of models[1]. We focused on the accuracy of models in the study, therefore, several parameters including IOU, PPV, AUC and TPR were employed.

5) Apart from model in the study, other deep learning methods including classification and identification could also aid in the diagnosis of ESCC. Moreover, the segmentation models such as DeepLabV3 or PsPnet might also be helpful in the segmentation of ESCC pathological sections. On the other hand, we think a larger cohort might be required for these models and only briefly mentioned these methods in the manuscript.

6) The mucosa segmentation model could aid in the tumor stage of ESCC, and tumor segmentation models could help pathologists quickly screen slides and look for suspicious areas. Moreover, previous studies have also demonstrated systems detecting neoplasia with endoscopic images and help clinicians determine the biopsy site, while the segmentation models could facilitate the diagnosis of biopsy specimen. The combination of multiple deep learning segmentation systems could assist the guidance of diagnosis and clinical decision, especially the feasibility of minimal invasive surgery (page 14, line 281-289).