Peer Review File

Article information: https://dx.doi.org/10.21037/tcr-23-964

Round 1

Reviewer Comments

I thank the authors for their comprehensive review of DP and AI. However, I do somehow miss the objective/message from the review. What do you want to say? There is multiple AI models? DP is efficient?. Further I think that this review is actually two reviews, and should be divided into two:

- One part is about AI (different models explained, different papers with AI reviewed). AI in clinical application.
- One part is about DP (workflow efficiency, TAT etc), also including a number of papers.
Otherwise, what has "Enemble learning" to do with TAT efficiency? Are there some models that promise more workflow efficiency than others? No.
- Interesting would be a review of which model types are in clinics nowadays

In my opinion, better separate the two topics.

Corresponding Author's Response: This review article was separated into two articles as suggested by the editors. The first part is the deep learning in digital pathology technical overview, and the second review will be the potential items needed for business development of digital pathology and DL healthcare workflows and real-world clinical use cases.
The article is intended to give healthcare providers, physicians, and business administrators an overview of what it takes to build this technology infrastructure into their hospital systems and the current successful clinical use cases in other hospital systems.

Further:
- Figure 2: (shouldn't this be a table?): Nice is the comparison of training and test data sets. But the metrics are not comparable at all, since the tasks are completely missing. Segmentation? Classiciation? Subtyping? Detection? Which cancer type? Which staining type?
Corresponding Author's response: correct these models are not comparable, it's on overview of the potential clinical use cases.

- Figure 3: seems to be out of context? What is that framework? There are many keywords, but what are these? What is LIS staff? What is QA staff? Is this missing? Is this required? What does it do?

Corresponding Authors' Response: Figure 1 corrected to Table 1. Figure 3 removed and will be placed in a separate clinical and business application of deep learning review. LIS means laboratory information systems. QA means quality assurance.

Minor points:
- Page 5 top, numbers are missing ("x patients...")
- Page 11 bottom, sections are empty
Corresponding Author's Response: Minor points corrected.

Overall, with the review makes a comprehensive (many papers involved), but somewhat sloppy and poorly written impression (inconsistent wordings, sometimes WSI, sometimes WSI's, sometimes sentences without verbs, ...). Many sentences are just put out, without a proper conclusion, or at least I have the impression there is a lot of data in the review, but little meaning. I think, this can be nicely addressed by separating the paper into two.
Further, the list of AI models seems more like a glossary than a review. What is the conclusion for the AI review?

Corresponding Author's response: The first review article (after the editors suggested the separation of sections) represents a deep learning technical review; most journal articles published to date represent the deep learning technical aspect, not the clinical application or the business use cases.

Conclusion for the first review article **Digital pathology and deep learning are emerging technologies that can improve clinical workflows and patient care. However, most published articles represent the pre-clinical deployment and model development phase. Future studies are needed to validate the real-world post-deployment production of deep learning.**

Round 2

Reviewer A

The authors provide an overview article of a rapidly growing field. Nicely, they used a systematic approach to identify relevant articles dealing with AI based whole slide scans. Because the field is rapidly advancing I do not want to propose big changes. However the authors miss some important fields likewise the quantification with AI trained alghorithms like NN192 and the latest publications using this. These should be mentioned in the discussion and referenced.

Response: NN192 algorithm summarized on page 15.

Reviewer B

The paper is well written and logically organized. Although great evidence supports digital adoption, real world data show still a limited adoption. Please discuss the evolution, the technical issues and how these have been faced (quote PMID: 37654717). Methods are described in detail with consistent results.

Response: DP evolution placed into the introduction section

Reviewer C

This narrative review aims to learn more about the performance of deep learning (DL) algorithms coupled with digital pathology (DP) regarding cancer. It is an interesting and perhaps necessary choice to go with a narrative account of ongoing studies about DL and DP. This is because these machine learning solutions are still in an early phase of development for pathology diagnosis. The authors of this review have done a good job of uncovering and listing many different features of DL and how they apply to pahtology analysis. However, the reason for my decision that the manuscript needs major revisions rests with the fragmented way that the authors present the topic and rationale of the study (introduction), methods, results, and discussion. All in all, in every section, there is a lack of coherence and organization of the topics. I will briefly identify the major problems and, where possible, suggest an approach to improving the manuscript.

Introduction: There is no connection between how DL relates to cancer diagnosis and improving healthcare costs and equity. Perhaps it is better to start with DL and DP in this section and then make a more logical connection with how DL and DP improve costs (some argue they don't) and how it all relates to healthcare equity.

Response: edited completed and updated the introduction section

Methods: There are few details about all the elements of PICO that are necessary for any type of review, but the major problem here is the comparison, i.e., model performance. While in some cases, machine learning applications used for diagnosis are very difficult to compare with human performance, it should be at least noted why such a comparison is difficult (or impossible). The problem here starts with the P (population in Cochrane guidelines, problem used by the authors). The moment where P is properly defined as 'human samples for cancer diagnosis' and not just 'cancer detection' like in the manuscript, it opens the possibility to compare it with a previous diagnosis by a pathologist, as samples for studies might be stored and previously diagnosed. Another important issue in the methods is the search string that is missing.

As a peer reviewer, I tend to question the validity of the article selection if I don't have at least a search string for one of the databases.

Response: edits completed and updated the method section.

Results: This section is probably the most fragmented. It starts with Table 1. Despite being a comprehensive overview of each study and its characteristics, it is necessary to have an explanation why the table columns are organized in that way and what the columns tell us about the forthcoming narrative of the results. The latter is just a list of the different features of DL used with DP and what they achieve. Often, there is no topic sentence or definition for the feature (see, for example, classic workflows or multiple-instance learning). The elements of the list of DL features or advantages are not coherent and do not follow clear logic. Why is there the topic 'Neural Image Compression' followed by 'Weakly Supervised Learning'? My suggestion for the whole results section is to group these fragmented topics into more major topics, e.g., group 'learning' together, and explain why you organized the topics this way. It is also necessary to highlight or make more visible what the reader can learn from this organizing principle.

Discussion: Similarly to the previous section, the discussion is fragmented and does not follow a clear logic for how it is organized. Moreover, there are too many topics discussed here. The idea of this section is to focus on the major insights from the results and discuss them at length, rather than presenting many features of DL and DP and discussing them in just one paragraph.

Response: It is difficult to compare each DL model study, because of the differences in data sets, models and metrics used, this is the reason why we chose to discuss the technical challenges faced during the DL development cycle

Finally, I believe that the manuscript could be improved and contribute to our knowledge about machine learning innovations in pathology. As I emphasized a few times, it is a matter of better organizing the findings and topics to be discussed, as well as more details and transparency about methods.

Round 3

Reviewer A

General Comments.
The review paper aims to discuss digital pathology and deep learning in the realm of cancer. The paper starts by presenting some facts about cancer, the deaths caused by it, and the worldwide costs incurred on cancer. The authors touched upon terms such as health equity, DP, TP, and DL. The main text starts with the PICO inclusions/exclusions

of the literature and jumps to various DL models and techniques used in the DP. After categorizing the short-listed literature in various DL headings, the paper gives some challenges and opportunities on the topic.
The paper is poorly written, with many typos, grammatical errors, incoherent sentences, out-of-context references, and semantic inconsistencies. It gives the reader a sense that the authors did not proofread the manuscript before submission.
The references were also not verified because many of them are wrongly cited and do not qualify the statements they are written with. I have given the line-wise details of such instances below. The paper is semantically incoherent and does not carry a smooth flow which is necessary in any academic literature.
The introduction and opportunities/ challenges sections do not conform to the main text given in the literature. For example, the Introduction and Challenges sections talk about TP and DL development lifecycle whereas the main text has no work related to these topics. Also, the Introduction talks about equity, bias, and ethics, but the main paper does not touch upon these concepts. Instead, the Challenges and Opportunities section gives some insight into some of the papers on these topics. Also, the introduction and abstract indicate that the paper will cover cancer detection, operational efficiency, costs, and patient care, whereas the Model Types section talks about cancer detection only on line 121. The same holds true for the selected papers in the manuscript's main text. All of these hiccups give a feeling that the Intro & Challenges sections were written separately from the main text leading to the sudden interruptions in the flow of thoughts.
The references are also poorly written, without considering any standard format. The paper requires a major review.

Line-wise Comments.
• Different font sizes (lines 20, 31, 47, 52-54, and many others). I lost track of so many typos. Response: Corrected
• Line 27: Unclear context of the sentence, “Several challenges arose during the model development that needed careful consideration.” Response: Corrected
• Line 29: Please introduce this DL lifecycle in one sentence. Response: Added
• Line 38: The cited reference [1] does not qualify the statement, “Cancer remains one of the leading causes of morbidity and mortality worldwide” Response: Citation corrected
• Line 42-43: “Public health systems worldwide spend $200 billion on cancer- related costs.” Please cite a reference to qualify this statement. Response: Citation corrected
• Line 43-44: Wrong reference to this statement, “By 2030, there will be 19.3 million new cancer cases and 10 million cancer deaths [3].” Response: Citation corrected
• Line 50-51: Wrong name of the agency given, “The Joint Commission and Institute for Healthcare Improvement.” Response: These are two agencies, The Joint Commission and the Institute for Healthcare Improvement
• Line 52: wrong use of the term, “analog” Removed
• Line 55-56: Please cite a reference to qualify this statement, “In 1986, Ronald Weinstein, MD, coined the term TP.” Response: Citation added
• Line 71-72: “digital image management systems and cloud-based platforms”, for

example? Response: reworded
• Line 73: $2,045.9 million or billion? Response: million
• Line 74: Please write full for the abbreviation CAGR. Response: written
• Line 74: [1][2] wrong references to the given statement. Response: corrected
• Line 77: Please define and explain the DL development lifecycle. Response: added
• Line 83: Why the Biorxiv was not considered in collection of databases? Response: It was, just forgot to add it into the methods section.
• Line 88: Typo Response: Corrected
• Line 94: Keywords are not required here, as already given below the abstract. Response: removed
• Line 118: Please avoid future tense in the manuscript. Response: Corrected
• Please use Vancouver referencing style as per TCR instructions.
• Reference numbering doesn't conform to the sequence in which they appear in the text (e.g., refs 13-17, 76 appear before refs 8-12 on line 57 and 65).
• Headings and subheadings are unclear because there are no numbering to the heading titles and the font size/types are the same. For example, Lines 117, 133, 131 and so on have the same fonts for section headings. Whereas reading the text gives one a sense that line 131 is a sub-heading of line 122.
• Line 122: DL-based DP workflows are not classical, instead they are an advanced version of the DP pipelines. Classical DP and Computational Pathology workflows involve mathematical and probabilistic modeling, bioinformatics/ biostatistics, and machine learning methods. Response: Corrected
• Line 135: "as each patch can capture unique features or characteristics that might be lost in a broader analysis." This is a wrong statement, learning on each patch may capture unique features, whereas the patch themselves are part of the bigger picture and may or may not contain the region of interest for tumor analysis. Response: Corrected
• Line 148: "but only the bag label is known". Response: Corrected
• Line 151: "minimum of 10,000 slides is recommended." Please cite the source. Response: Completed
• Line 154: Please define EPL before introducing the term. Response: Corrected
• Line 159: CNNs abbreviation already defined. Response: Corrected
• Line 161: DTFD wrongly mentioned. Response: Corrected
• Line 164: please define WSISA. Response: Corrected
• Line 168: please rewrite the sentence. Response: Rewritten
• Line 169: Wrong statement, the cited work is not related to Breast Cancer. Response: Corrected
• Line 172: Missing citation for saMIL. Response: Corrected
• Line 174: Wrong full for MIMS. Response: Removed
• Line 184: Wrong citation for MT-MIL. Response: Corrected
• Line 197: Wrong statement, the cited work is not related to Colon Cancer. Response: Corrected
• Line 201: [55] is not related to medical data, hence wrong citation Response: Removed
• Line 201: "methods construct a graph from the image using graph convolutional networks" this is not how graphs are constructed in Graphs-based methods. Response:

removed and corrected
• Line 231: [18] is not an MIL paper. Response: Removed
• Line 249, 253: Typos Response: Corrected
• Line 265: [18] is not a medical data-related paper. Response: removed and added digital pathology papers.
• Line 270: what is TURP? Response: transurethral resection of the prostate
• Line 277: [20], [41] and [53] are not ensemble-based works. (Need works)
• Line 302: [48] doesn't use NN192, consider removing it from this section. Response: Removed
• Line 304: Wrong conclusions from [52], the paper does not present any model development workflow. Response: Removed
• Line 328: "Conformal prediction in healthcare can increase patient safety" is a wrong conclusion from the cited paper. The cited paper talks about the patient safety of AI systems and not patient safety in general. Response: Correct
• Line 331: References in the heading are not acceptable in academic publications.
• Line 333: typo Response: Corrected
• Line 331: Challenges and Opportunities talk about DL development lifecycle but this term hasn't been explained anywhere in the paper.
• Line 360: How is this a challenge/ opportunity? Response: In MILs, local patches have limited visual context; opportunities arise to improve these computational inefficiencies, for example, by using Graph-based methods.
• Line 371: "DL models are often trained on datasets different from those tested on." This is a wrong sweeping statement, DL models are usually task-specific and the training-evaluation is done for the same task and the same type of data. Response: Removed, adjusted
• Line 392: Without any mention in the main paper about TP, the term has been used here without the context.
• Line 397: [69] is wrongly cited as it is not related to bias or equity. Response: removed
• Line 399: Please define the abbreviations JCAHO and IHI. Response: Corrected
• Line 405: Please cite and mention examples of such methods to qualify the statement. Response: methods listed and citations provide
• Lines 406-413: Terms like TP, compression, latency, and communications do not conform to the material provided in the main paper. Write full for EMPAIA. Response: Full form provided. ecosystem for pathology diagnostics with AI assistance
• Line 447: Wrong referencing [76] is not related to pre-trained models. Response: removed
• Reference 28, 23, 22: Wrong citation Response: Corrected
• All references are in the incorrect format. Response: Corrected
• Figure 1 has blurred text.