**IEEE SPMB 2023: Review Results**

**Submission Type:** Abstract

**Abstract No. A005:** Big Data Resources for Digital Pathology

**Score:** 9/10

**Summary:**

It seems the abstract is too big and should be converted to a paper instead.

Reviewers have several suggestions for improving the abstract as well as for future research.

**Reviewer 1:**

*Section A*: Outline the abstract.

In this abstract, the authors introduce two new datasets that are made publicly available: more than 14,000 newly released images from the Fox Chase Cancer Center and around 100,000 images from the Temple University Digital Pathology Corpus (TUDP). They have also presented the changes in the organization of the dataset to make it more unified. The images in these datasets are digitized using the protocols used for TUDP. To facilitate analysis using machine learning algorithms, each image has been assigned a tumor grade or Gleason Score. They also give details of the naming convention for the filenames.

*Section B*: Highlight the contribution of the abstract.

This work will have a huge impact on the field. They have publicly released two large datasets consisting of diagnostic images. These images are labeled paving path for machine learning tools for analysis in the future. The authors have also invested a lot of effort in coming up with an informative naming convention which will further enable ease of data analysis.

*Section C*: Please state your recommendations and why.

The authors have elegantly conveyed all the information in this abstract.

*Section D*: Please state ways to improve the abstract.

*Section E*: Miscellaneous comments.

**Reviewer 2:**

*Section A*: Outline the abstract.

The paper introduces their recently released digitized data, FCDP, and the changes they made to TUDP inorder to align it with FCDP, making it compatible with machine learning systems.The FCDP corpora encompasses more than 13,000 subjects, annotated in various ways.The existing TUDP dataset issue was its insufficiency; it was only partially annotated and the filenames did not adhere to the authors’ desired standard. Subsequently, the data from FCCC was released. The authors proceeded to digitize this FCCC data, labeling it as FCDP. This corpus, being considerably larger and possessing more standardized filenames, was then used as a basis to implement parallel changes to TUDP.

*Section B*: Highlight the contribution of the abstract.

This paper provides a set of metadata for each slide, including slide and specimen identification numbers, tissue information, tumor details, staining method, and grading information.The paper also addresses cases where grading information couldn’t be determined. It provides an approach to categorize such cases as ”unknown.”

*Section C*: Please state your recommendations and why.

I would suggest a more logical structure for the paper (SPMB paper structure format). Instead of describing FCDP in Part A and TUDP in Part B, followed by a summary in Part C, it would be more effective to integrate the information.It seems that the labels of whether the tumors are malignant or benign are needed in this study because toa large extent, the treatment of the tumor depends on knowing this.

*Section D*: Please state ways to improve the abstract.

The paper mentions cases where the grade couldn’t be determined and how these were categorized as ”un-known”. It would be helpful to elaborate on how these cases might impact the usability of the dataset. Thepaper briefly mentions the possibility of augmenting FCDP with TUDP. Actually, I could not find where they explained this augmenting process. Providing more details on how this augmentation can be done and the potential benefits would be valuable

*Section E*: Miscellaneous comments.

The first sentence would benefit from stronger support than what is provided by Reference1 and its manner of citation.2.Page 2 of 19 paragraph 2, line 3: The error in the sentence is the repetition of ”shown in shown in.” It should be corrected to: ”The data was classified into the same four cate-gories shown in Table 1 using the information found in columns AF (“Block Level TissueHistology”) and AG (“Gradeclin Desc”).”3.page 2, table 1, While they mentioned the existence of seven types of tumors (four cate-gorized as high-risk and three as general), Table 1 does not reflect these seven types, and it raises questions about why the unknown type is denoted as ’X,’ corresponding to 10 in Roman numerals.4. Page 2 of 19 paragraph 2, line 5: the period shouldn’t be after “unknown”.5.Table 5 title style is different from other tables

**Reviewer 4:**

*Section A*: Outline the abstract.

The abstract discusses the introduction of a new corpus of digital pathology images, the Fox Chase Digital Pathology (FCDP) Corpus, containing 14,288 high-resolution images collected from the Fox Chase Cancer Center's Biosample Repository. The authors also mention the changes made to the Temple University Digital Pathology Corpus (TUDP) to organize these corpora in a unified framework. The abstract provides an overview of the data, its storage format, metadata, and classification criteria for machine learning. It also briefly mentions the TUDP Breast Tissue Corpus and its relationship to the FCDP data.

*Section B*: Highlight the contribution of the abstract.

The abstract highlights the significance of these datasets in the field of digital pathology, especially due to their large size and diverse representation of tissue types and staining methods. It also mentions the potential for future annotations and unsupervised learning experiments, which could further enhance the value of the datasets.

*Section C*: Please state your recommendations and why.

The abstract in the present form could be accepted for the conference with minor clarifications.

The authors introduce the release of two significant datasets in the field of digital pathology. This work is useful for researchers interested in utilizing these datasets for machine learning experiments or other research purposes. The explanation of the datasets' sources and ethical considerations is commendable, demonstrating transparency and adherence to ethical standards.

*Section D*: Please state ways to improve the abstract.

* The abstract mentions an imbalance in the data, but it could provide more details or statistics to quantify this imbalance.
* More information about the imaging process and any preprocessing steps performed on the data would be useful.
* The abstract introduces the classification criteria for tumors based on grading systems. Is it verified by an oncology researcher?
* Are there any potential limitations or challenges associated with the datasets to provide a balanced view for potential users?

*Section E*: Miscellaneous comments.

Providing more context and potential applications for these directions would be helpful. This could be provided if there is a room with respect to the word limit.

Overall, this abstract introduces valuable digital pathology datasets and provides essential details for potential users.