

## Peer Review File

---

Automated multi-scale computational pathotyping (AMSCP) of inflamed synovial tissue



**Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

## Reviewers' Comments:

### Reviewer #1:

#### Remarks to the Author:

The authors have performed extensive analysis of the application of several machine learning methods including random forest, gradient boosting and neural networks to histology slides in RA (rheumatoid arthritis) for the purpose of developing predictive models to identify pathotypes and histological features on slides from synovial tissue. An interesting aspect of their approach is the use of slides from mice with experimental arthritis as training data. Overall the analysis is detailed and multiple approaches have been tried. The machine learning has been attempted at a deep level and the authors are to be congratulated on an extensive piece of work. Overall there are important results suggesting that their ML models can identify histological features in human synovial biopsy slides in RA.

However, there are some methodological points which need to be clarified especially revolving around measurement of performance, use of CV and it is critical that there is clarity around the use of hold-out test sets. I was also slightly disappointed that the main predictive model is not made public with code on github or as a python package (or whatever language the model is coded in). This is important for reproducibility/validation in future cohorts. The authors state that scripts will be made available on reasonable request, but this makes it hard for reviewers and potential future readers to scrutinise the code of the models. One major issue is that the training data is entirely H&E and that immunohistochemistry (IHC) with CD markers was not used to confirm cell type.

#### Major points:

1. What were the total number of whole slide images from mice used for training?
2. How many 512x512 pixel patches were used in total for training?
3. Supplementary Fig. 1 & 2 are helpful, but would be aided by adding the total and subtotal number of patches.
4. How many images and patches were used for testing performance?
5. This is partially defined in line 414 (171 slides). How many patches?
6. What software package was used for the GBDT? Gradient-boosted tree models generally definitely need CV to tune their parameters especially the stopping parameter. Please provide much more detail about how the GBDT was tuned in the methods (line 290). CV was used: were the same CV folds used to measure performance? If so, this needs to be declared and it should be acknowledged that this will be highly likely to inflate performance measurements unless separate hold out test set or nested CV have been employed. If these hold out techniques have been used, they must be described (what proportion of hold out etc). Some information appears on line 443, but it is still unclear if separate hold out sets were used for performance measurement or tuning and performance were done on the same CV hold out folds. What was stratified (line 443)? Stratification is not mentioned in line 291.
7. Please comment on the imbalanced class size for testing when applied to 3 pathotypes (only 5 in pauci-immune group).
8. mIOU is defined on page 648. Please define this abbreviation at the top of p. 9 line 209. It is the abbreviation is expanded on line 230. Add a sentence explaining what mIOU means (or define it properly), as this article will be read by a general audience.
9. Ditto fwIOU (frequency weighted IOU) which is first encountered on line 371 but the abbreviation is not defined. Please full define in the methods.
10. Figure 1. Define how many replicates of analysis were performed for each model in 1A. Were analyses repeated using bootstrap or CV? Were different hold out sets used, or was it the same hold-out set for all model fitting? If bootstrap or CV was used to define training and test sets, were these folds kept constant across all model fitting, or were they randomly assigned for each repeat

for each model? Line 385: this is a critical result. Overall there is insufficient information (number of slides, number of patches used for training, proportion of data held out or whether separate test set was employed) to explain how models were trained and tested at this juncture. If not using CV/ bootstrap but using a single separate test set, then how many times was the test set queried overall? This is important as otherwise all that is being honed is how best to achieve results on this particular test set.

11. Or: are the results quoted from lines 385 to 406 simply the result of training and testing on the same data? If true, it needs to be clarified. If repeated hold-outs are used, there must be much more granularity about the details of the process.

12. Line 196 how were the RF parameters determined? Were these determined by trial and error/ empirically or by CV. If CV was employed as part of the results lines 385-406, then performance can only be measured in an unbiased way if nested CV is used. Was nested CV employed? If not, this needs to be stated.

13. Fig. 2A correlation plot is driven by 2 outliers and only 9 points appear to be present (?). Why were so few samples in the validation set? Was it not possible to have a larger validation set? It is a problem that the data are not more evenly distributed. Correlation is ok, but RMSE between true result and predicted result would be better as r-squared can be erroneous if there is a consistent absolute deviation from ground truth. Please report RMSE as well as r squared (Fig. 2A + line 413). The fact that the gradient is not 1 probably needs some explaining. Also, the intercept in the model is reported in the figure as 107041. But that means the line of best fit should meet the y axis at  $1 \times 10^6$  when it meets closer to 0. Please check the formula for y is correct.

14. Fig. 2 legend is discursive and sounds more like results text. Words like "Interestingly" should be removed from the figure legend (although they could be moved to the results). It should simply explain the figures and their origin in more detail including statistical tests. For example, r-squared is presumably Pearson here not spearman? What statistical tests are used in 2C? How is area measured in C? This is incredibly important. Does this show area as measured by the old gold standard? Or area as measured by segmentation and machine learning (presumably by the UNET++ model as mentioned in the title)? Please clarify. Does each dot represent a mouse or a joint? What placebo was used? This information needs to be provided in the figure legend. Fig. 2D please fully explain these images and the colour segmentation of their overlay. What do the overlay colours refer to? What was the original underlying slide stained with?

15. Fig. 3D boxplot axis fonts are too small. Please enlarge.

16. Fig. 4 legend. Please document what statistical methods were used. I assume Spearman's rho is used in Fig. 4D, E, and Pearson r-squared is used in 4F. These need to be declared. Line 690, the authors have shown a correlation between log lymphocyte counts. The word "log" is missing.

17. Why have Spearman and Pearson correlation analysis been used differently in different plots in Fig 4 and 6? Please justify why different correlation methods were used for different plots. Why not use Pearson applied to log lymphocyte measurements? They appear to be strongly linear once log scale is applied.

18. Fig. 6 legend. Please document statistical methods again. Again correlation with log lymphocyte counts is depicted in 6B, but the word "log" is missing from the legend.

19. Line 438: please add additional information: how many slides? I am assuming it is 19 slides as mentioned in line 290 in the methods? The subimages passed to the GBDT are presumably segmented by cell. It would help to make the results clearer if some additional essential information from the methods was included as otherwise the reader constantly has to go back and forth between results and methods.

20. Line 475. Please explain random sampling in this context. If you retrain the model on a smaller subset of the data, are you reusing model parameters and selected/filtered features learnt from the whole dataset? If so, this is not a valid way to reassess performance.

21. If 5-fold CV is being used to measure performance (line 478) then is the same CV used for tuning models? If so, this would bias performance measurement. Nested CV would be necessary.

22. Fig 6C, please add some high resolution images showing the cell types being correctly identified.

23. One issue is that these cell types are not identified by immunohistochemistry (IHC) with specific markers. They are being identified by morphology only which has limits. In particular it would be difficult to be sure of the distinction between some of these cell types using H&E alone. On the other hand the method applied is machine learning applied to H&E. It is interesting that the ML method once trained, does well, but it really needs to be properly benchmarked against IHC stained slides. This caveat/ limitation must be mentioned in the discussion. The authors do make it clear in line 562 that they used H&E stained slides. But the lack of training the model against a gold standard using IHC with different CD markers should be added.

24. How did the authors handle imbalanced data? Were any balancing techniques employed?

Minor points:

1. Please provide a list of all software packages used, their primary installation language, source repository and add version numbers for all of these.
2. Which version of Python was used (line 350)?

Reviewer #2:

Remarks to the Author:

Bell et al. describe a deep learning method that can perform image segmentation and cell type classification on rheumatoid arthritis histopathology images. The results are interesting and I really did want to be enthusiastic about them, but for me, they were occluded by serious shortcomings in the description and openness about the approaches used. In particular, my biggest concerns, and the key issues that prevent me from recommending publication in its current state, are the lack of transparency owing to a lack of detail in the methods section, lack of any code or data provided to the reviewers, and a data sharing policy that seems very behind the times.

There are significant gaps in the machine learning details and this work is not reproducible with the amount of data given. In particular:

- "We then performed color and shape feature extraction of each super-pixel." How?
  - Almost no details were given on the random forest classifier. What hyperparameters were used?
  - A pre-trained EfficientNet-B5 was used for the UNET++ encoder (why?) but as far as I can tell, we have no information on the decoder or any of the other UNET++ parameters so we really don't know what was used here.
  - Line 224: "To calculate performance of each type of overlap we used to Validation set with mixed training." I don't know what you mean here.
  - "All model hyperparameter optimization was performed on the validation set...." How?
  - The "detailed workflow for the ImageJ/FIJI analysis" is not detailed enough to be reproducible or for us to really understand what the authors did. E.g., "Characteristics of neighbouring cells at several different distance ranges were used to include tissue level context into cell type classification." If we imagine ourselves sitting down at a computer to try to reproduce this, sentences like this really get to the heart of the issue I'm describing: How can we code something as vague as "several different distance ranges" and expect to reproduce what has been done?
  - "We developed novel custom feature extraction methods..." I don't know which feature extraction methods you're referring to - the ones in my first point? There doesn't seem to be any detail, so how do we know that they're novel? Why are they novel? Why did they have to be novel?
- While the methods section needs revision to include more detail and to be written in a much more precise way, all of this speaks to a greater issue: No methods section can ever be detailed enough to truly describe what software is doing. This is why source code is shared and released, but none was provided, so we are unable to assess the quality of the code and whether it actually does what the authors say it does. The data sharing policy for this paper is also disappointing: "All original

data and analysis scripts will be provided upon reasonable request." I'm very surprised that this is even allowed, but if we assume that it is, I'm puzzled by this choice as it seems to act against the authors' own interests. Version controlled open source repositories have been available for over a decade – why not just put all of the source code and analysis scripts in a repository under an open source license so that users can scrutinize and access it whenever they want? This is a method which the authors surely want to be widely used and adopted so it can generate impact, so why create additional barriers to entry for users to access their tool? If you do then share the code, what license would that be under and what are users allowed to do with it? What happens if the corresponding author moves institution and the Email address provided is no longer valid? It's also not clear to me if the authors are actually willing to share everything (including models and weights) as describing these as "analysis scripts" would be strange. For the public, this approach to data sharing seems outdated and bad for open science. For the reviewers, we haven't been able to scrutinize a key aspect of this work.

Taken together, it really is impossible to know what has been done here as there are serious issues with transparency and reproducibility. It's a shame that what could have and should have been interesting work was occluded by such fundamental issues, but I would encourage the authors to address the above comments in full and I oppose publication until they are adequately addressed.

Minor:

- Line 222: "to removes" --> "to remove"

## REVIEWER COMMENTS

Dear Editor and Reviewers,

We are overwhelmingly grateful for this opportunity to respond to the Reviewers' critiques and wholeheartedly agree with their main criticism. We believe that their criticism can be broken down into four main categories: 1) lack of providing the scripts for critical review, 2) a modest level of naiveté about classification training structure (nested CV vs regular CV), 3) to include IHC validation of the cell class model, and 4) some oversight in how we describe our methods. To address these four main points, we have re-written major portions of the manuscript and: 1) created a github page with associated Google drive links for all the scripts and models (Link: [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath)) which is public and licensed under the CC-BY-NC-ND 4.0 license, 2) built new cell classification models using a nested CV structure with a large grid search for GBDT parameters and re-calculated the data for Figures 3-6, 3) acquired 15 cases from the set of 60 human RA biopsies with adjacent section IHC stained for relevant CD and other markers and performed cell type validation, and 4) re-doubled our effort in describing our methods. Our point-by-point responses to specific comments are detailed below.

Reviewer #1 (Remarks to the Author):

The authors have performed extensive analysis of the application of several machine learning methods including random forest, gradient boosting and neural networks to histology slides in RA (rheumatoid arthritis) for the purpose of developing predictive models to identify pathotypes and histological features on slides from synovial tissue. An interesting aspect of their approach is the use of slides from mice with experimental arthritis as training data. Overall the analysis is detailed and multiple approaches have been tried. The machine learning has been attempted at a deep level and the authors are to be congratulated on an extensive piece of work. Overall there are important results suggesting that their ML models can identify histological features in human synovial biopsy slides in RA.

However, there are some methodological points which need to be clarified especially revolving around measurement of performance, use of CV and it is critical that there is clarity around the use of hold-out test sets. I was also slightly disappointed that the main predictive model is not made public with code on github or as a python package (or whatever language the model is coded in). This is important for reproducibility/validation in future cohorts. The authors state that scripts will be made available on reasonable request, but this makes it hard for reviewers and potential future readers to scrutinise the code of the models. One major issue is that the training data is entirely H&E and that immunohistochemistry (IHC) with CD markers was not used to confirm cell type.

Major points:

1. What were the total number of whole slide images from mice used for training? [See Response below](#)
2. How many 512x512 pixel patches were used in total for training? [See Response below](#)

3. Supplementary Fig. 1 & 2 are helpful, but would be aided by adding the total and subtotal number of patches. [See Response below](#)

4. How many images and patches were used for testing performance? [See Response below](#)

5. This is partially defined in line 414 (171 slides). How many patches?

**Response:** Points 1-5 are addressed in the new table include in Supp. Fig 1 copied below.

|                                            | Mixed Training Strategy |                    |             |  | Single Batch Training |                    |             |  | Grand Totals |
|--------------------------------------------|-------------------------|--------------------|-------------|--|-----------------------|--------------------|-------------|--|--------------|
|                                            | Trainin<br>g Set        | Validatio<br>n Set | Test<br>Set |  | Trainin<br>g Set      | Validatio<br>n Set | Test<br>Set |  |              |
| Total<br>Knees<br>Whole<br>Slide<br>Images | 62                      | 16                 | 16          |  | 42                    | 20                 | 32          |  | 94           |
| Total<br>Patches<br>*                      | 80,487                  | 16,101             | 21,598      |  | 56,730                | 25,650             | 35,806      |  | 118,186      |

### Supplemental Figure 1. Training, Testing and Validation with batch effect study design.

A) Our H&E sections for training testing and internal validation were acquired from two batches (n=94). All of these slides were fully annotated for 11 tissue types. Batch A were sections obtained from the TNF transgenic model of inflammatory arthritis and wildtype (WT) littermates that were sectioned and stained at the University of Rochester Medical Center and previously published in Bell et al 2019<sup>a</sup>, and Bell et al 2019<sup>b</sup>. Batch B were sections from the Zymosan induced arthritis model that were section and stained at the Hospital for Special Surgery.

B) For the Mixed Training Strategy (Left), we first performed stratified random sampling to separate the slides into a Training, Test, and Validation set, stratifying on disease type (healthy vs diseased) and batch (Batch A vs Batch B). For the Single Batch Training Strategy (Right), we also performed a sampling which placed all of Batch B in the Validation set and combined the Batch A slides from the Test and Validation sets into the Test set. The numbers represent the number of slides in each set.

C) Table summarizing the total number of slides, and number of 512 px by 512 px patches in each of these sets. \* the number of patches was calculated using the 60% overlap patch generation method that demonstrated the best performance.

6. What software package was used for the GBDT? Gradient-boosted tree models generally

definitely need CV to tune their parameters especially the stopping parameter. Please provide much more detail about how the GBDT was tuned in the methods (line 290). CV was used: were the same CV folds used to measure performance? If so, this needs to be declared and it should be acknowledged that this is highly likely to inflate performance measurements unless separate hold out test set or nested CV have been employed. If these hold out techniques have been used, they must be described (what proportion of hold out etc). Some information appears on line 443, but it is still unclear if separate hold out sets were used for performance measurement or tuning and performance were done on the same CV hold out folds. What was stratified (line 443)? Stratification is not mentioned in line 291.

**Response:** Based on this comment and comments from the 2<sup>nd</sup> Reviewer, we have now utilized a nested 5-fold CV approach for all of the classification results. We have updated the Methods accordingly (Line 349-360 and copied below).

“A total of 4,712 cells were annotated for mouse cell type classification from seven different classes (Supplemental Fig. 8). Cells were labeled from a total of 19 different slides. A GBDT model was trained for cell type classification (XGBoost (<https://xgboost.readthedocs.io/en/stable/>) implemented within the Scikit-learn package (<https://scikit-learn.org/stable/>)), using stratified nested 5-fold cross validation with grid search to select the best models. In order to minimize the influence of annotations from any one slide, we enforced an even sampling method to ensure approximately equal amounts of cells from each slide appeared in all folds (`sklearn.model_selection.StratifiedKFold`). To tune the parameters of the GBDT, we performed a grid search on the inner CV for learning rate = [0.05, 0.1, 0.2], `colsample_bytree` = [0.6, 0.8, 1.0], `subsample` = [0.25, 0.5], `max_depth` = [6, 12], `n_estimators` = [10, 100, 200, 400], `gamma` = [0, 0.1, 0.3], and `min_child_weight` = [1, 5, 10]. To evaluate model performance, F1 statistics were calculated as the average of the 5 external folds and then the best performing models were used to infer cell type in two different biological settings: (1) to identify cell composition changes across different disease severities on the remaining cells (>300,000) on the 19 slides, and (2) identify differences between male and female mice in the context of disease progression in a held-out dataset [18].”

7. Please comment on the imbalanced class size for testing when applied to 3 pathotypes (only 5 in pauci-immune group).

**Response:** The authors were concerned about this exact point and thus when abstracting any modeling to the subject level, we avoided any inference about the pauci-immune pathotype (ie saying we can identify a pauci immune pathotype even though high fibroblast % would be pretty good indicators in many cases) except to plot the number of cells identified by the classification model. When annotating the slides and training the model, we: 1) attempted to annotate all types of cells on all pathotypes, although for example, there are very few lymphocytes on pauci-immune cases, and 2) we employed a stratified sampling CV to enforce that the sampling method draws approximately equal numbers of each cell type from each case. We feel that we have appropriately walked the line of utilizing our dataset to its maximum without over fitting or over concluding.

8. mIOU is defined on page 648. Please define this abbreviation at the top of p. 9 line 209. It is

the abbreviation is expanded on line 230. Add a sentence explaining what mIOU means (or define it properly), as this article will be read by a general audience.

**Response:** We have updated the text as requested (Line 191).

9. Ditto fwIOU (frequency weighted IOU) which is first encountered on line 371 but the abbreviation is not defined. Please full define in the methods.

**Response:** We have updated the text as requested (Line 191- 192).

10. Figure 1. Define how many replicates of analysis were performed for each model in 1A. Were analyses repeated using bootstrap or CV? Were different hold out sets used, or was it the same hold-out set for all model fitting? If bootstrap or CV was used to define training and test sets, were these folds kept constant across all model fitting, or were they randomly assigned for each repeat for each model? Line 385: this is a critical result. Overall there is insufficient information (number of slides, number of patches used for training, proportion of data held out or whether separate test set was employed) to explain how models were trained and tested at this juncture. If not using CV/ bootstrap but using a single separate test set, then how many times was the test set queried overall? This is important as otherwise all that is being honed is how best to achieve results on this particular test set.

See Response to comments #10 and 11 below

11. Or: are the results quoted from lines 385 to 406 simply the result of training and testing on the same data? If true, it needs to be clarified. If repeated hold-outs are used, there must be much more granularity about the details of the process.

**Response** to Comments 10 and 11: We apologize for the lack of clarity in our segmentation training regime. We have re-written parts of the methods section (Line 174-257) to clarify the procedure. The main modified section of the method that responds to these comments is copied below.

“Initially, a stratified random sampling method was used to randomize the 94 annotated WSIs into a training, validation, and test split, using 70%, 15%, and 15% for each split, respectively (3 splits total, **Supplemental Figure 1B**). We stratified by batch (e.g. staining and site differences) and disease type (e.g. healthy and disease) to ensure even allocation of data variation into each set. Randomization occurred at the slide level, as opposed to the patch level, to ensure no data leakage across splits. During our initial qualitative explorations of models and hyperparameters (detailed below), we only utilized training and validation sets; and calculated the Dice score on the validation set to measure performance. Dice score was calculated as

$$Dice = \frac{2 * intersection}{union + intersation}$$

These initial experiments included variations in SLIC feature extraction, model selection (UNET vs UNET++), efficient-net backbone size (B0, B2, B5), loss metrics (weight loss vs unweighted), and learning rate parameters as described below. Once these items were tuned, quantitative experiments (details below) were performed varying the tissue segmentation number (7, 9, 10, or 11), image augmentation (None, Low, Medium, or High), or patch overlap percentage (0%, 50%, 66%). These experiments were performed by training the models with the Training and Validation sets, freezing the models' weights, and then inferring segmentation

on the Test set. These inferences were then compared to the ground truth labels to calculate the mean Intersection over Union (mIOU) or the frequency weighted mIOU (fwIOU).

In addition, to assess how site-specific differences in histology slides may impact model performance [28], we performed a single batch training method while varying the image augmentation style. In this training strategy, Batch A (n=62) was used for the training (45%) and validation (20%) sets, and Batch B (n=32, 35%) was used as the held-out test set. These experiments were performed to assess if image augmentation could overcome batch related staining differences that are seen in the real world."

12. Line 196 how were the RF parameters determined? Were these determined by trial and error/ empirically or by CV. If CV was employed as part of the results lines 385-406, then performance can only be measured in an unbiased way if nested CV is used. Was nested CV employed? If not, this needs to be stated.

**Response:** We have rewritten methods section (Lines 222-240). Because QuPath provides no scripting implementation of ML learning training methods (i.e. CV, data set splitting, sampling etc.) and is mostly a GUI to interact directly with the image data, all implementation of our methods were done manually. Despite this, a primary reason why we decided to implement at least one model fully in QuPath was because it is the gold standard digital pathology software and wanted to demonstrate its qualities (i.e. it is decent at 7 class segmentation), but also show its limitations. However, this restricted our ability to algorithmically explore any hyper parameters. Therefore, we used the default implementation in QuPath. We did manually explore some SLIC feature extraction methods to arrive at the input features which has now been detailed and copied below.

"To compare segmentation performance using both conventional machine learning and deep learning methods, we tested two different model architectures. A Random Forest (RF) model implemented in QuPath with OpenCV [30] and an U-Net++ [31] deep learning model implemented in PyTorch [32]. We qualitatively assessed both pixel level segmentation and super-pixel level segmentation in QuPath and determined Super-Pixel segmentation performed better. To generate super-pixels, we applied a Simple Linear Iterative Clustering algorithm ( $\sigma = 5$ , spacing = 20  $\mu\text{m}$ , Max Iterations = 1, Regularization = 0.01) in which each over segmented area was considered a super-pixel. We qualitatively assessed SLIC feature extraction variations within the training and validation set and choose to extract RGB values, estimated Hematoxylin, Eosin and residual stain means, standard deviations, min, max, and median values from each SLIC super pixel. We also calculated the Haralick value using a distance of 1 and bin of 32. We next calculated the average features of super-pixels within 40 and 80  $\mu\text{ms}$ . These features were then used in the RF, which was implemented fully in QuPath using their "Train Object Classifier" GUI with max depth = 20, min sample count = 10, Active variable count = 0, maximum trees = 50 and termination epsilon = 0. Four models were built, one for each of the 7 class, 9 class, 10 class and 11 class segmentation tasks. After the models were trained, they were used to infer on the Test set and performance metrics calculated."

13. Fig. 2A correlation plot is driven by 2 outliers and only 9 points appear to be present (?). Why were so few samples in the validation set? Was it not possible to have a larger validation set? It is a problem that the data are not more evenly distributed. Correlation is ok, but RMSE

between true result and predicted result would be better as r-squared can be erroneous if there is a consistent absolute deviation from ground truth. Please report RMSE as well as r squared (Fig. 2A + line 413). The fact that the gradient is not 1 probably needs some explaining. Also, the intercept in the model is reported in the figure as 107041. But that means the line of best fit should meet the y axis at  $1 \times 10^6$  when it meets closer to 0. Please check the formula for y is correct.

**Response:** As the Test set (n=16) was randomly selected a priori, we find it acceptable that these results are driven by the apparent “outliers”, as the data set was randomly sampled from a much larger set. Only n=9 of the test set were from Bell et al., 2019, which had corresponding hand-drawn synovial area measurements. This set is also from an experiment that had a wide range of disease that was cross-sectionally sampled (i.e., 3 months of age, 4 months of age and 5.5 months of age), of which 2 were healthy, 6 had mild disease and 1 had severe disease that contributes to the scatter of the data (i.e. 2 very low samples and 1 very high). The graph below shows the results when all samples from Bell et al., 2019 are included regardless if they were randomized into the Training, Validation, or Test sets. We have now included the RMSE to the analysis and checked that the equation is correct. Slope less than 1 and Y intercept >0 likely mean that in healthy samples (low synovial tissue areas) the model is over estimating a bit, which makes sense as in healthy tissue deciding what is synovial capsule vs synovium can be difficult or inconsistent to measure; and at large synovial areas, or severely diseased tissues, the model is underestimating as the synovium may be invading other tissues which confuse the predictions. We have added text in the results to point out this limitation (Line 529-532)

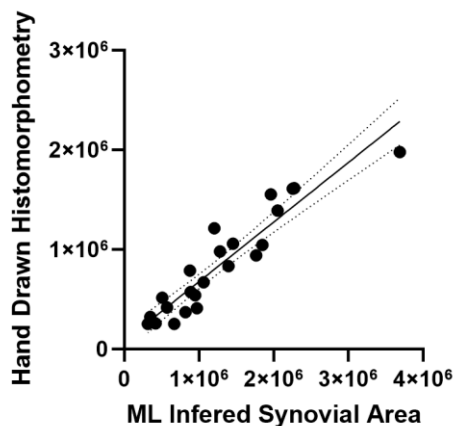


**Figure 1. Pearson correlation of all samples from Bell et al 2019 regardless of whether they were randomized into the Training, Validation, or Test sets.** This graph represents all of the data that is consistent between the current work and Bell et al., 2019, which has historical hand drawn histomorphometry. The graph presented in Fig 2A is a subset of this data, including only the Test set.

14. Fig. 2 legend is discursive and sounds more like results text. Words like ???Interestingly??? should be removed from the figure legend (although they could be moved to the results). It should simply explain the figures and their origin in more detail including statistical tests. For example, r-squared is presumably Pearson here not spearman? What statistical tests are used in 2C? How is area measured in C? This is incredibly important. Does this show area as measured by the old gold standard? Or area as measured by segmentation and machine

learning (presumably by the UNET++ model as mentioned in the title)? Please clarify. Does each dot represent a mouse or a joint? What placebo was used? This information needs to be provided in the figure legend. Fig. 2D please fully explain these images and the colour segmentation of their overlay. What do the overlay colours refer to? What was the original underlying slide stained with?

**Response:** We have re-written all figure legends to include more descriptive methods and with little to no interpretation. Figure 2 legend is copied below to answer the Reviewer's specific questions.

**“Figure 2. The fine-tuned UNET++ model measures treatment response in the TNF-Tg with Anti-TNF therapy.**

(A) Fine-tuned UNET++ inferred synovial area in the held-out test set compared to historical hand drawn histomorphometry (n=9, Pearson's Correlation).

(B) Anti-TNF therapy Study design. Six-month-old WT and TNF-Tg mice were used as controls. Eight-month-old WT or TNF-Tg mice were treated with Anti-TNF therapy or Placebo control for 6 weeks (weeks post treatment, wpt). Both left and right knees were collected and 2-3 histologic levels per knee were analyzed (n = Slides / Knees).

(C) Using the fine tuned UNET++ model we inferred tissue area of the Synovium (Left), Trabecular Bone (Middle) and Cartilage (Right) for 6mo WT and TNF Controls as well as Placebo (Irrelevant IgG) treated WT, Placebo treated TNF-Tg and Anti-TNF treated TNF-Tg. Each dot represents one knee (average of 2-3 histologic levels), Box and Whisker plots are construct by showing the Min, 25<sup>th</sup> %tile, Median, 75<sup>th</sup> %tile and Max, One-Way ANOVA with Tukey's Post-hoc Test, \*p<0.05, \*\*p<0.01, \*\*\*p<0.001, \*\*\*\*p<0.0001

(D) Representative images (2x magnification) of 6mo WT and TNF Controls as well as Placebo treated WT, Placebo treated TNF-Tg and Anti-TNF treated TNF-Tg with predicted tissue overlay.”

15. Fig. 3D boxplot axis fonts are too small. Please enlarge.

**Response:** We have updated as suggested.

16. Fig. 4 legend. Please document what statistical methods were used. I assume Spearman's rho is used in Fig. 4D, E, and Pearson r-squared is used in 4F. These need to be declared. Line 690, the authors have shown a correlation between log lymphocyte counts. The word log is missing.

**Response:** We have updated the legend to address the Reviewer's comments. We did not correlate on log transformed data, we performed a Spearman's Correlation on non-normal data (that happens to be log-normal). Figure legend is copied Below:

**“Figure 4. Computational pathology modeling recapitulates the sexual dimorphism of TNF-Tg inflammatory arthritis.**

An independent set of slides from the training slides were used to validate the cell type prediction model (n=3-6 per group; Male and Female; WT and TNF; 3-, 4- and 5.5-month-old).

Tissue segmentation was first performed to segment the synovium and then cell type predictions were performed within the synovium.

(A) Lymphocytes predictions counts.

(B) Synovial Lining Cell prediction counts

(C) Fibroblast predictions counts.

(A-C) Each dot represents one mouse, Box and Whisker plots are construct by showing the Min, 25<sup>th</sup> %tile, Median, 75<sup>th</sup> %tile and Max, Two-Way ANOVA with Tukey's Post-hoc Test, \*p<0.05, \*\*p<0.01, \*\*\*\*p<0.0001, Differences only shown between the female and male TNF-Tg mice.

(D) Lymphocyte predictions compare to the synovial inflammatory score as previously quantified in Bell et al, 2019. Spearman's correlation, TNF-Tg mice only (n=28).

(E) Synovial Lining cell counts compared to the pannus invasion score as previously quantified in Bell et al, 2019. Spearman's correlation, TNF-Tg mice only (n=28).

(F) Total cell counts in the synovium compared to the cell area as previously quantified in Bell et al, 2019, Pearson's correlation, TNF-Tg mice only (n=28)."

17. Why have Spearman and Pearson correlation analysis been used differently in different plots in Fig 4 and 6? Please justify why different correlation methods were used for different plots. Why not use Pearson applied to log lymphocyte measurements? They appear to be strongly linear once log scale is applied.

**Response:** Thanks for your comment. As we have explained in the methods section (Line 455-459) "For all continuous variables, a Shapiro-Wilks Normality test was performed to assess normality. If the test determined the specific distribution to be non-normal, the equivalent non-parametric test was utilized to test for significance or correlation."

Thus, 4D and E failed normality and F did not. In addition, we supply a detail statistical statement "Supplemental Statistical Information Pertaining to Data Presented in Figures" to further describe the statistical methods.

18. Fig. 6 legend. Please document statistical methods again. Again correlation with log lymphocyte counts is depicted in 6B, but the word ???log??? is missing from the legend.

**Response:** We have updated Figure 6 legend. Please note that the offending graph has been moved to Figure 5, which has also been re-written (see below). While we do not perform any transformations on the data and thus the units are not log, we do prefer to choose graph axis scales which appropriately shows all of the data, which is log. As such, we now have included notes to point out that the scales are log.

**Figure 5. Cell type modeling correctly classifies synovial stromal and immune cells in RA synovial biopsies.**

(A) A subset of cells from 13 RA synovial biopsies were annotated (n=2,341) using an active learning strategy. After nuclei detection and custom feature extraction from each cell, a gradient

boosted decision tree was trained using a parameter grid search with a nested, stratified, 5-fold cross-validation training strategy. The F1 scores ( $M \pm SD$ ) of the five folds for each cell class are presented with the overall weighted F1 of  $0.85 \pm 0.01$  ( $M \pm SD$ ).

(B) The confusion matrix from the most performant model demonstrates the typical misclassification in this dataset (data is cell counts). Stromal cells can be mistaken for other stromal cells (vascular endothelial cells, synovial lining cells, and fibroblast) and lymphocytes can be mistaken for plasma cells. F : Fibroblast, L : Lymphoid, M/H : Macrophage/Histocyte, PC : Plasma Cell, S/C : Stromal/Connective Cell, SLC : Synovial Lining Cell; VEC : Vascular Endothelial Cell.

(C-D) The most performant model was used to predict the cell type of the remaining cells from all RA synovial biopsies (n=60)

(C) Adjacent sections to the H&E-stained slides were stained with either CD3 (T-Cells), CD20 (B-Cells) and CD138 (Plasma Cells) or CLIC5 (Synovial Lining), CD3 (T-Cells), CD68 (Macrophages), and CD34 (Vascular Endothelial Cells) (n=10) . Representative images of plasma cells ( $C^i$ ), lymphocytes ( $C^{ii}$ ), synovial lining cells ( $C^{iii}$ ) and vascular endothelial cells ( $C^{iv}$ ) with the original H&E in the left column, prediction overlays in the middle, and adjacent slide IF in the right column. Immunostains and magnification are denoted within the image.

(D) Correlation of machine learning predictions with quantitative histomorphometry of IF+ cells from adjacent sections of n=10 RA synovial biopsy pieces. Top: CD138+ cells vs ML Predictions of Plasms cells; Middle: CD3+ and CD20+ cells vs ML Predictions of Lymphocytes; Spearman's Correlations (n=10). Please note the log scale.

(E) Correlation of machine learning predictions of lymphocytes (as a percent of total cells) vs the Krenn Inflammation Score, Spearman's Correlations (n=60). Please note the log scale on the y-axis.

19. Line 438: please add additional information: how many slides? I am assuming it is 19 slides as mentioned in line 290 in the methods? The subimages passed to the GBDT are presumably segmented by cell. It would help to make the results clearer if some additional essential information from the methods was included as otherwise the reader constantly has to go back and forth between results and methods.

**Response:** We have updated the description in the results section (Line 558-569) as below.

“We annotated a total of 4,712 cells across three stages of murine inflammatory arthritis in the TNF-Tg mice (n=6 healthy, n=8 mild disease, and n=5 severe, **Supplemental Figure 8**) to build a cell type classification model that could recognize cells from various disease stages. Cell nuclei were first segmented with HoverNet [39], nuclei boundaries passed to ImageJ where stain deconvoluted color features of both the nuclei and cytoplasm as well as nuclei shape parameters were calculated. We next passed these data to our custom feature leveraging our novel insight that cell type predictive modeling dramatically improves with neighborhood features (Supplemental Table 3) inspired by adjacent work tackling a different classification task[43]. These neighborhood features include standard statistical measure of neighboring cells within a radius to the parent cell like hematoxylin mean or kurtosis, whether the parent cell was

located in a dense region of other cells, and the shape (convex hull) of the cells within a radius to the parent cell (Supplemental Table 2). ”

20. Line 475. Please explain random sampling in this context. If you retrain the model on a smaller subset of the data, are you reusing model parameters and selected/filtered features learnt from the whole dataset? If so, this is not a valid way to reassess performance.

**Response:** During the active learning experiments, we explored 4 methods to select cells for the next iteration of annotation selection (Line 312): “smallest margin uncertainty [44], least confidence uncertainty [45] and entropy-based uncertainty [44]” and random. Random means we randomly chose 5% of the data to add to the training set before the next round of training. The full descriptions (Line 378-399) are copied below:

“Human annotation was the time-consuming step for the cell type classification pipeline. Therefore, we applied the active learning strategy to improve the annotation efficiency for cell annotation of human samples. To develop this strategy, we tested a proof-of-concept active learning strategy using labeled data from the mouse H&E slides (**Supplemental Figure 9**). Active learning is an iterative process that consists of three main steps, (1) annotation, (2) model training, and (3) sample selection for further annotations. Its goal is to select the samples that can lead to the largest model performance improvement when adding to the training data after annotation. To validate the strategy, 100 different rounds of 5-fold cross-validation were performed. Average F1-scores were reported for each class and a macro-F1 score was additionally reported. 25 runs of 5-fold cross validation were removed due to cells from a single class not being present in both the training and testing sets. For the training dataset for each split, 5% of cells were first randomly selected as the first set of cells selected as being labeled annotated. Subsequently, the GBDT classifier was trained using this randomly selected data. Several different metrics for determining cells for annotation and subsequent model finetuning, including smallest margin uncertainty [44], least confidence uncertainty [45] and entropy-based uncertainty [44] were assessed. The top 5% of cells were added to the training dataset and the cycle of model training and evaluation and new cell annotations continued until the entire training dataset was used. A random selection of cells after shuffling was also tested to compare model performance to the various active learning strategies. The package modal [46] was leveraged in our implementation. Mean and 95% confidence intervals are reports for each subset across the 75 different runs of 5-fold cross validation.”

21. If 5-fold CV is being used to measure performance (line 478) then is the same CV used for tuning models? If so, this would bias performance measurement. Nested CV would be necessary.

**Response:** We thank the Reviewer for this suggestion. We have now altered all of our cell classification training procedures to be a nested 5-fold CV with grid search. Given this, all the data in Figures 3-6 are now changed. The descriptions (Line 350-361) are copied below:

“In order to minimize the influence of annotations from any one slide, we enforced an even sampling method to ensure approximately equal amounts of cells from each slide appeared in all folds (sklearn.model\_selection.StratifiedKFold). To tune the parameters of the GBDT, we performed a grid search (sklearn.model\_selection.GridSearchCV) on the inner CV for learning rate = [0.05, 0.1, 0.2], colsample\_bytree = [0.6, 0.8, 1.0], subsample = [0.25, 0.5], max\_depth = [6, 12], n\_estimators = [10, 100, 200, 400], gamma = [0, 0.1, 0.3], and min\_child\_weight = [1, 5, 10].”

22. Fig 6C, please add some high resolution images showing the cell types being correctly identified.

**Response:** High resolution images have been added in both Fig 5 and 6.

23. One issue is that these cell types are not identified by immunohistochemistry (IHC) with specific markers. They are being identified by morphology only which has limits. In particular it would be difficult to be sure of the distinction between some of these cell types using H&E alone. On the other hand the method applied is machine learning applied to H&E. It is interesting that the ML method once trained, does well, but it really needs to be properly benchmarked against IHC stained slides. This caveat/ limitation must be mentioned in the discussion. The authors do make it clear in line 562 that they used H&E stained slides. But the lack of training the model against a gold standard using IHC with different CD markers should be added.

**Response:** We completely agree with the Reviewer that morphology and H&E feature have limits. We have acknowledged this limitation in the Discussion (Line 759-769). In addition, we have included new data from immune stained biopsies in Fig 5 to support the validation of the ML predictions. These results appear on line 631-661, and are copied below. However, we do not want to use this as training data for 2 reasons: 1) we fundamentally wanted to develop a model that was purely derived from H&E features, and 2) in our current dataset, we only have adjacent slides which do not guarantee to contain exactly the same cell.

“To further validate our human RA cell typing model, we acquired adjacent immunostained sections (Lymphoid, n = 4; Diffuse, n = 4; Pauci-Immune, n = 2) for lymphoid (CD3, CD20 and CD138) and stromal-immune (CD3, CD68, CLIC5 and CD34) markers to qualitatively and quantitatively assess for cell type validity. Our qualitative analysis revealed a remarkable spatial alignment of machine learning predicted plasma cells with CD138+ cells (**Figure 5C<sup>i</sup>**, green outline middle panel vs green IF right panel) and lymphocytes with CD3+ or CD20+ cells (**Figure 5C<sup>ii</sup>**, blue outline middle panel vs red/white IF right panel), synovial lining cells with CLIC5+ cells (**Figure 5C<sup>iii</sup>**, light blue outline middle panel vs yellow IF right panel) and vascular endothelial cells with CD34+ cells (**Figure 5C<sup>iv</sup>**, grey outline middle panel vs red IF right panel) in many of the sections. We also observed some alignment of macrophage-histiocytes predictions with CD68+ cells (**Figure 5C<sup>iii</sup>**, purple outline middle panel vs green IF right panel) but these observations were much less consistent suggesting that either the model does not predict these cells well in a broader setting or there is not a lot of similarity of a H&E define histiocyte and a CD68+ macrophage or both. We were unable to acquire immunostains to validate our fibroblast and stromal-connective cell classes. Lymphocyte, and in particular B-cells, infiltration into the tissues is an important pathologic finding to discriminate RA pathotypes [2, 11, 52] and classify disease severity [23]. Therefore, to quantitatively validate our lymphocyte and plasma cell machine learning predictions vs fiducial protein cell markers on adjacent slides, we performed thresholding histomorphometry analysis to count the number of either DAPI+, CD138+ or DAPI+, CD3+ or DAPI+, CD20+ cells within regions of interest. There is an excellent correlation between our machine learning predictions of plasma cells vs DAPI+, CD138+ ( $\rho = 0.86$ ,  $p < 0.002$ , **Figure 5D**, top), and lymphocytes with vs DAPI+, CD3+ or DAPI+, CD20+ cells ( $\rho = 0.97$ ,  $p < 0.0001$ , **Figure 5D**, bottom). Krenn inflammation score is an important, expert

derived histopathology score driven mainly by the number of lymphocytes that are present in the whole synovial biopsy specimen. Thus, we calculated the percent of machine learning predicted lymphocytes to total cells and correlated with the Krenn inflammation score in all samples (n=60). This revealed a high degree of correlation between the predictions and pathologist scores ( $\rho = 0.79$ ,  $p < 0.0001$ , **Figure 5E**). Taken together, this analysis demonstrates we have an excellent synovial cell type prediction model that faithfully predicts lymphoid cells which mimic a clinical pathologist score.”

24. How did the authors handle imbalanced data? Were any balancing techniques employed?

**Response:** During the segmentation task (deep learning), we investigated a weighted loss metric to account for the large imbalance between some of the tissue annotation areas. However, we did not observe performance improvements (sometimes even worse performance) other than the unweighted loss metric. For the cell classification task, we did not feel that there was enough imbalance to explore alternate methods as each cell type is between 7-20% frequencies in human and 6-15% except for lining cells which are a diverse group of cells at ~30% of the data.

Minor points:

1. Please provide a list of all software packages used, their primary installation language, source repository and add version numbers for all of these.

**Response:** All packages and version are detailed in the python environment spec-file for both segmentation and cell classification in the github repository (Link: [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath), Username: arthritis-histopath, Password: NatComm123). We have also updated the methods sections (Line 465-474) which is copied below.

*“Software, Hardware, Dataset and Code Availability*

All machine learning techniques were performed in Python (3.8.1) as described above. Primary machine learning libraries include PyTorch (1.8.0), segmentation\_models\_pytorch (0.1.3) [33] Sklearn (scikit-learn, 1.3.2), and xgboost (2.0.2). All analysis scripts and models are provided at [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath). Full environment dependencies are in the spec-files in the Segmentation or Cell Classification folders. Deep learning segmentation training was performed on four Nvidia V100’s GPUs with 16 gb of RAM in parallel. Segmentation inference was performed on either a Nvidia 3070 or 3090. Cell type classification was performed on an Nvidia 3070. Original image data will be provided upon reasonable request in accordance with institutional data use sharing agreements.”

2. Which version of Python was used (line 350)?

**Response:** Python 3.8.10

Reviewer #2 (Remarks to the Author):

Bell et al. describe a deep learning method that can perform image segmentation and cell type classification on rheumatoid arthritis histopathology images. The results are interesting and I really did want to be enthusiastic about them, but for me, they were occluded by serious shortcomings in the description and openness about the approaches used. In particular, my biggest concerns, and the key issues that prevent me from recommending publication in its current state, are the lack of transparency owing to a lack of detail in the methods section, lack of any code or data provided to the reviewers, and a data sharing policy that seems very behind the times.

There are significant gaps in the machine learning details and this work is not reproducible with the amount of data given. In particular:

- ???We then performed color and shape feature extraction of each super-pixel.??? How?

**Response:** We have added links to the github repository (Link: [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath), with all of the scripting details reported. We have also updated the text to be more descriptive. Copied here:

“To generate super-pixels, we applied a Simple Linear Iterative Clustering algorithm ( $\sigma = 5$ , spacing = 20  $\mu\text{m}$ , Max Iterations = 1, Regularization = 0.01) in which each over segmented area was considered a super-pixel. We qualitatively assessed SLIC feature extraction variations and choose to extract RGB, estimated Hematoxylin, Eosin and residual stain means, standard deviations, min, max, and median values from each SLIC super pixel. We also calculated the Haralick value using a distance of 1 and bin of 32. We next calculated the average features of super-pixels within 40 and 80  $\mu\text{ms}$ .”

- Almost no details were given on the random forest classifier. What hyperparameters were used?

**Response:** We are grateful to the Reviewer for this question. As we have responded to the comments from the first Reviewer, RF was implemented completely in the GUI provided by QuPath. All hyperparameters are default implementation. We have copied the relevant text (Line 234-240) as below.

“These features were then used in the RF, which was implemented fully in QuPath using their “Train Object Classifier” GUI with max depth = 20, min sample count = 10, Active variable count = 0, maximum trees = 50 and termination epsilon – 0. Four models were built, one for each of the 7 class, 9 class, 10 class and 11 class segmentation tasks. After the models were trained, they were used to infer on the Test set and performance metrics calculated.”

- A pre-trained EfficientNet-B5 was used for the UNET++ encoder (why?) but as far as I can tell, we have no information on the decoder or any of the other UNET++ parameters so we really don't know what was used here.

**Response:** We utilized an ImageNet pre-trained encoder for better convergence quality. For computational efficiency (i.e. GPU VRAM loading) consideration, we used an efficient architecture, EfficientNet-B5, but left the decoder unchanged from the UNET++ native architecture. We have updated the methods (Line 244-250) to reflect this additional detail, copied here:

“For our deep learning pipeline, we utilized the segmentation\_models\_pytorch [33] package for the UNET++ implementation with an ImageNet pre-trained EfficientNet-B5 backbone for the encoder to improve training time and computational efficiency [34]. The decoder was left unchanged from the native UNET++ architecture except for the final layer which was changed to match the number of tissue types being segmented. We also initially explored using the UNET architecture and other efficient net backbones, however UNET++ with a B5 backbone was found to be most performant.”

- Line 224: ???To calculate performance of each type of overlap we used to Validation set with mixed training.??? I don???t know what you mean here.

**Response:** As detailed in Supplemental Figure 1, the validation set from the mixed training strategy (as opposed to the Test set from the single batch training) is a specific set of slides. We use this set during tile overlap tested to measure model performance. We have updated the methods section (Line 174-191) to reflect this as below.

“Initially, a stratified random sampling method was used to randomize the 94 annotated WSIs into a training, validation, and test split, using 70%, 15%, and 15% for each split, respectively (3 splits total, **Supplemental Figure 1B**). We stratified by batch (e.g. staining and site differences) and disease type (e.g. healthy and disease) to ensure even allocation of data variation into each set. Randomization occurred at the slide level, as opposed to the patch level, to ensure no data leakage across splits. During our initial qualitative explorations of models and hyperparameters (detailed below), we only utilized training and validation sets; and calculated the Dice score on the validation set to measure performance. Dice score was calculated as

$$Dice = \frac{2 * intersection}{union + intersction}$$

These initial experiments included variations in SLIC feature extraction, model selection (UNET vs UNET++), efficient-net backbone size (B0, B2, B5), loss metrics (weight loss vs unweighted), and learning rate parameters as described below. Once these items were tuned, quantitative experiments (details below) were performed varying the tissue segmentation number (7, 9, 10, or 11), image augmentation (None, Low, Medium, or High), or patch overlap percentage (0%, 50%, 66%). These experiments were performed by training the models with the Training and Validation sets, freezing the models’ weights, and then inferring segmentation on the Test set. These inferences were then compared to the ground truth labels to calculate the mean Intersection over Union (mIOU) or the frequency weighted mIOU (fwIOU).”

- ???All model hyperparameter optimization was performed on the validation set???.??? How?

**Response:** We apologize for the lack of clarity. Segmentation model design choices were informed by looking at the Dice score of the validation set (which is the internal biased set). We have added detailed descriptions on this procedure (Line 179-185) and copied below:

“During our initial qualitative explorations of models and hyperparameters (detailed below), we only utilized training and validation sets; and calculated the Dice score on the validation set to measure performance. Dice score was calculated as

$$Dice = \frac{2 * intersection}{union + intersction}$$

These initial experiments included variations in SLIC feature extraction, model selection (UNET vs UNET++), efficient-net backbone size (B0, B2, B5), loss metrics (weight loss vs unweighted), and learning rate parameters as described below.”

- The ???detailed workflow for the ImageJ/FIJI analysis??? is not detailed enough to be reproducible or for us to really understand what the authors did. E.g., ???Characteristics of neighbouring cells at several different distance ranges were used to include tissue level context into cell type classification.??? If we imagine ourselves sitting down at a computer to try to reproduce this, sentences like this really get to the heart of the issue I???m describing: How can we code something as vague as ???several different distance ranges??? and expect to reproduce what has been done?

**Response:** We apologize for the lack of clarity. The FIJI script and subsequent python script are uploaded to our github page (Link: [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath)). We have added addition text in the methods to further describe this procedure (Line 321-332, copied below) and included new Supplemental Tables 1 and 2 that detail all 856 features used in classification.

“These features are call cell intrinsic features (**Supplemental Table 1**). Neighborhood characteristics of cells at several different distance ranges (150 px, and 300 px) were used to include local cell and tissue level context into cell type classification (**Supplemental Table 2**) [43]. Simply, a neighborhood is determined as all the cells within a certain distance to the parent cell. These neighborhood characteristics included the average, standard deviation, skew, kurtosis, Z-score, interquartile range, standard error of the mean and entropy the cell intrinsic features of cells in the neighborhood. We also calculated shape characteristics of the neighborhood including the average distance of the neighboring cells, the linear correlation coefficient of the cells within the neighborhood, a string of up to 30 cells linear correlation coefficient, the strait line distance of a string of up to 30 cells, a scored measure of density of the cells and the number of cells within the distance measure. A final total of 854 features were extracted for the downstream analysis.”

- ???We developed novel custom feature extraction methods?????? I don???t know which feature extraction methods you???re referring to - the ones in my first point? There doesn???t seem to be any detail, so how do we know that they???re novel? Why are they novel? Why did they have to be novel?

While the methods section needs revision to include more detail and to be written in a much more precise way, all of this speaks to a greater issue: No methods section can ever be detailed enough to truly describe what software is doing. This is why source code is shared and released, but none was provided, so we are unable to assess the quality of the code and whether it actually does what the authors say it does. The data sharing policy for this paper is also disappointing: ???All original data and analysis scripts will be provided upon reasonable request.??? I???m very surprised that this is even allowed, but if we assume that it is, I???m puzzled by this choice as it seems to act against the authors??? own interests. Version controlled open source repositories have been available for over a decade ??? why not just put all of the source code and analysis scripts in a repository under an open source license so that users can scrutinize and access it whenever they want? This is a method which the authors surely want to be widely used and adopted so it can generate impact, so why create additional barriers to entry for users to access their tool? If you do then share the code, what license would that be under and what are users allowed to do with it? What happens if the corresponding author moves institution and the Email address provided is no longer valid? It???s also not clear to me if the authors are actually willing to share everything (including models and weights) as describing these as ???analysis scripts??? would be strange. For the public, this approach to

data sharing seems outdated and bad for open science. For the reviewers, we haven't been able to scrutinize a key aspect of this work.

Taken together, it really is impossible to know what has been done here as there are serious issues with transparency and reproducibility. It's a shame that what could have and should have been interesting work was occluded by such fundamental issues, but I would encourage the authors to address the above comments in full and I oppose publication until they are adequately addressed.

**Response:** We are most grateful for these comments and cannot agree more on the importance and necessity of promoting open data science. We have created a github repository with all source codes and documentation available (Link: [https://github.com/rdbell3/Arthritis\\_HistoPath](https://github.com/rdbell3/Arthritis_HistoPath)) in this revision, and we encourage all researchers to use our code under the CC-BY-NC-ND 4.0 license. We will continuously maintain and update the repository based on the feedbacks we receive.

As far as the authors can tell from a detailed review of the literature, we are the first to utilize neighborhood features in a cell classification task, with only 1 other work (Wang et al, Ref. # 43) using them for any prediction task. As such, we do feel this is a novel approach that needs to be highlighted. We have added discussion points to more clearly describe the impact of our work (Line 720-730), copied below:

"In designing our feature extraction pipeline, we were inspired by Wang et al [43] who used cell intrinsic features, contextual features about surrounding nuclei, density features and spatial arrangement features generated from a H&E tumor biopsy in a graph neural net framework to predict if patients would benefit from check-point inhibitor therapy. We extend this feature modeling framework to include both hematoxylin and eosin specific features, as well as cytoplasm features with statistical calculations of neighborhood features that empirically show improved that cell classification performance (Cell Intrinsic Features:  $0.75 \pm 0.01$  vs Cell Intrinsic Features plus 150 px distance features:  $0.86 \pm 0.02$  vs All Features:  $0.88 \pm 0.03$ , Supplemental Table 3 and Figure 3). To the authors knowledge, our work is the first to utilizes this type of framework in a cell type classification task."

Minor:

- Line 222: ???to removes??? --> ???to remove???

**Response:** We have changed this.

## Reviewers' Comments:

### Reviewer #1:

#### Remarks to the Author:

The authors have significantly improved the article by addressing the individual points which I previously raised. Critically they have now provided source code in a recognised public repository (GitHub). They have provided an example dataset for model inference, which aids reproducibility.

They have also moved to using nested CV which reduces the bias in performance measurement when the dataset is limited. They did use data augmentation, so there might be a potential information leak here (it's not clear whether data augmentation was performed on training data only). But overall, given that the performance was similar between nested CV and non-nested (single) CV, I suspect that the augmentation is necessary for image recognition training, but probably does not inflate performance estimation by much. True (real world) estimate of performance would only be achievable with a completely independent validation dataset. The authors should add words to this effect in the Discussion, i.e. something along the lines of "a recognised limitation is that the performance of the model in the real world still requires formal validation in an independent cohort".

#### Minor comments:

Please add the version of CUDA used in the main model training on the Nvidia V100 GPUs and the version of QuPath used.

### Reviewer #2:

#### Remarks to the Author:

My main critique of this manuscript was to do with code sharing, data availability, and reproducibility of the methods. I believe the authors have now adequately addressed my concerns, I commend them for doing so, and congratulate them on a nice piece of work. I only have one minor comment on the GitHub repository (see below) but this is only a suggestion and I'll leave its implementation up to editorial discretion.

Dear Editor and Reviewers,

We are very grateful for the acceptance, in principal, of our manuscript. We have address all of the reviewer concerns and the editorial concerns with this resubmission. Below we have responded specifically to the minimal comments.

Thank you,

Richard Bell, PhD

Reviewer #1 (Remarks to the Author):

The authors have significantly improved the article by addressing the individual points which I previously raised. Critically they have now provided source code in a recognised public repository (GitHub). They have provided an example dataset for model inference, which aids reproducibility.

They have also moved to using nested CV which reduces the bias in performance measurement when the dataset is limited. They did use data augmentation, so there might be a potential information leak here (it's not clear whether data augmentation was performed on training data only). But overall, given that the performance was similar between nested CV and non-nested (single) CV, I suspect that the augmentation is necessary for image recognition training, but probably does not inflate performance estimation by much. True (real world) estimate of performance would only be achievable with a completely independent validation dataset. The authors should add words to this effect in the Discussion, i.e. something along the lines of "a recognised limitation is that the performance of the model in the real world still requires formal validation in an independent cohort".

Minor comments:

Please add the version of CUDA used in the main model training on the Nvidia V100 GPUs and the version of QuPath used.

Response: We have added the CUDA and Qupath versions to the methods section.

Reviewer #1 (Remarks on code availability):

I have not formally installed and run the code, as this could take an excessive length of time. I have reviewed the code and it appears to be sufficiently detailed and organised, although it could annotated in more depth. The authors provide sizeable example data (3 GB csv file) which is downloadable. I was able to download the csv file and inspect it, and it seems reasonable. Notably there are instructions for performing inference from the models. There is a spec file which includes sklearn dependency information (packages, version numbers). Overall this should probably be sufficient for the community to scrutinise the method and models.

Reviewer #2 (Remarks to the Author):

My main critique of this manuscript was to do with code sharing, data availability, and reproducibility of the methods. I believe the authors have now adequately addressed my concerns, I commend them for doing so, and congratulate them on a nice piece of work. I only have one minor comment on the GitHub repository (see below) but this is only a suggestion and I'll leave its implementation up to editorial discretion.

Reviewer #2 (Remarks on code availability):

The newly provided code in the GitHub repository seems appropriate, the instructions seem to have an appropriate level of detail, and the CC license is fine. While I don't have a problem with providing the segmentation models via Google Drive per say, I would suggest including these in the version control (unless they're prohibitively large). This will make it easier to fix users' problems in the future if models change over time.

Response: Unfortunately, github restricts the file size and our models are larger than that limit. We do not know of another alternative at the moment.