Journal of Pathology Informatics 14 (2023) 100333

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Analysis of cellularity in H&E-stained rat bone marrow tissue via deep learning

Smadar Shiffman [a], Edgar A. Rios Piedra [a], Adeyemi O. Adedeji [a], Catherine F. Ruff [a], Rachel N. Andrews [a], Paula Katavolos [a,b], Evan Liu [c], Ashley Forster [a,d], Jochen Brumm [e], Reina N. Fuji [a], Ruth Sullivan [a,*]

[a] Genentech Research and Early Development (gRED), Department of Safety Assessment, Genentech Inc., South San Francisco, USA
[b] Bristol Myers Squibb, New Brunswick, NJ 08901, USA
[c] Genentech Research and Early Development (gRED), Department of Development Sciences Informatics, Genentech Inc, South San Francisco, USA
[d] University of Pennsylvania School of Veterinary Medicine, Philadelphia, PA 19104, USA
[e] Genentech Research and Early Development (gRED), Department of Nonclinical Biostatistics, Genentech Inc, South San Francisco, USA

ARTICLE INFO

Keywords:
Cell quantification
Deep learning
Digital pathology
H&E rat bone marrow

ABSTRACT

Our objective was to develop an automated deep-learning-based method to evaluate cellularity in rat bone marrow hematoxylin and eosin whole slide images for preclinical safety assessment. We trained a shallow CNN for segmenting marrow, 2 Mask R-CNN models for segmenting megakaryocytes (MKCs), and small hematopoietic cells (SHCs), and a SegNet model for segmenting red blood cells. We incorporated the models into a pipeline that identifies and counts MKCs and SHCs in rat bone marrow. We compared cell segmentation and counts that our method generated to those that pathologists generated on 10 slides with a range of cell depletion levels from 10 studies. For SHCs, we compared cell counts that our method generated to counts generated by Cellpose and Stardist. The median Dice and object Dice scores for MKCs using our method vs pathologist consensus and the inter- and intra-pathologist variation were comparable, with overlapping first-third quartile ranges. For SHCs, the median scores were close, with first-third quartile ranges partially overlapping intra-pathologist variation. For SHCs, in comparison to Cellpose and Stardist, counts from our method were closer to pathologist counts, with a smaller 95% limits of agreement range. The performance of the bone marrow analysis pipeline supports its incorporation into routine use as an aid for hematotoxicity assessment by pathologists. The pipeline could help expedite hematotoxicity assessment in preclinical studies and consequently could expedite drug development. The method may enable meta-analysis of rat bone marrow characteristics from future and historical whole slide images and may generate new biological insights from cross-study comparisons.

Introduction

A critical step in pharmaceutical drug development is preclinical safety assessment, which involves evaluating the effects of candidate molecules in *in-vivo* animal models prior to human dosing in clinical trials. A major part of safety assessment is the microscopic evaluation of all major organ systems in these animals for signs of drug-related toxicity. A common drug-related finding is hematotoxicity, toxicity to the blood cells and blood-producing tissues. The bone marrow is the primary site for hematopoiesis, or blood cell production, making its evaluation crucial in preclinical safety studies prior to submission of Investigational New Drug (IND) applications.

Microscopic evaluation of the bone marrow is routinely performed on hematoxylin and eosin (H&E)-stained tissue sections by toxicologic pathologists in the preclinical setting, and includes assessment of the cellularity, or cell count, proportion of blood cell types, and blood cell maturation.[1–6] A common histopathological finding in the bone marrow is decreased cellularity of hematopoietic elements, which is based on the anatomic pathologist's manual visual assessment of bone marrow sections by microscope, and scored on a scale of severity, from minimal through mild, moderate, marked, to severe.[7] This manual visual assessment, on glass slides[4] or whole slide images (WSIs),[8] is typically performed at multiple magnifications to allow scrutiny of both tissue architecture and cellular morphology. Fig. 1 demonstrates some of the magnifications required for evaluating the cellularity in a whole slide image (WSI) of bone marrow from a rat with mild cell depletion. Manual assessment can be subjective,[7] laborious, and time-consuming,[9] and is prone to inter-pathologist variability.[10–12] In preclinical toxicity studies, pathologists compare bone marrow cellularity across multiple dose groups, often against a control group, and ensuring accuracy and consistency can be challenging in this subjective, qualitative approach.[1]

A variety of automated methods for analyzing WSIs have been developed that help pathologists obtain accurate and reproducible analysis

* Corresponding author.
E-mail address: sullivan.ruth@gene.com (R. Sullivan).

http://dx.doi.org/10.1016/j.jpi.2023.100333
Received 22 June 2023; Received in revised form 18 August 2023; Accepted 19 August 2023
Available online 25 August 2023

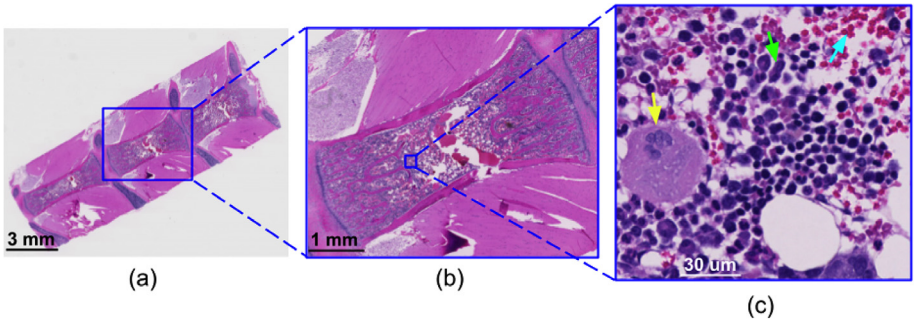

Fig. 1. Whole slide image of rat sternum (H&E), acquired at 40× magnification (pixel size 0.23 μm). (a) A section through the sternum. (b) A section through a sternebra. (c) A view through the marrow, with arrows pointing to examples of a megakaryocyte (MKC) in yellow, myeloid cells in lime green, and red blood cells (RBCs) in cyan.

results with considerably less effort compared to the standard evaluation methods.[6,7,11–13] To achieve high accuracy and robustness, methods for automated analysis of WSIs must address the challenges related to tissue characteristics, sample preparation, and imaging artifacts, all of which could impact the appearance of cells in WSIs.

Challenging tissue characteristics include a multitude of different types of cells[14] that are densely mixed and heterogeneously distributed,[15] touching and/or overlapping cells, potentially indistinct cell boundaries, and natural variations in the appearance of cells within cell classes.[16] For example, the myeloid cell appearances vary across different cell lineages and stages of development. Another challenge is the variation in the staining quality of the H&E slides that may result from differences in sample preparation and staining procedures. Further, underlying pathology can impact staining properties. For example, Fig. 2 shows variation in the tinctorial appearance of the small hematopoietic cells (SHCs)[1] and megakaryocytes (MKCs) across WSIs with increasing cell depletion.[2]

Typical artifacts in WSIs include defects in slide preparation such as tissue tears, folds, staining variability within and across laboratories, as well as blurriness and noise which are defects that can arise from slide scanning.[17–19] In current automated image analysis methods, it is common practice to remove artifacts where possible in a pre-processing step. Preprocessing addresses the variability in tissue characteristics by extracting explicitly or implicitly the important information in WSIs, such as color, cell morphology, nuclear orientation, texture, and spatial arrangement, and associating this information with objects of interest for analysis. Another approach to addressing the anticipated variability is selection of a large set of samples that exhibits the variability, or via data augmentation.[20]

We leveraged recent deep learning model architectures to develop a high-throughput screening pipeline for rapid, accurate, and robust quantification of cellularity in WSIs of bone marrow tissue for preclinical safety studies performed in rats. While we could have adapted existing models for analyzing histopathological WSIs as other groups did,[21–23] we anticipated that tailoring a pipeline specifically for rat histopathological WSIs would yield high accuracy, which is crucial for hematotoxicity evaluation. In this paper, we refer to the steps that the pipeline performed as *our method*. We describe the post-processing techniques that we developed to address detection and segmentation errors made by the deep learning models, which were limited by the small amount of data available for training, under-representation of samples with regions of interest (ROIs), and the difficulty in annotating cell ground truth (GT). We demonstrate that our method identified and segmented bone marrow objects on an evaluation dataset with an accuracy that was close to that of pathologists and out-performed 2 state-of-the-art publicly available deep-learning cell segmentation models.

Automated analysis of pathology images began in the early 1990s.[24] The introduction of WSIs enabled wide application of image processing techniques, such as smoothing, thresholding, morphological operators, and filters, to analyze histopathology WSIs, including human and animal bone marrow data.[7,12] WSIs allowed computational scientists to leverage machine learning methods to develop efficient and robust prediction algorithms with reduced dependency on tuning parameters via trial and error for a variety of tissues,[8,18] including human bone marrow.[25,26] Machine learning processes included extracting representations from sample image data and training models by comparing predictions based on these representations to ground truth. In the last decade, deep learning, a subset of machine learning, has enabled development of highly accurate, efficient, and generalizable image analysis models, which generally outperformed traditional methods as well as human performance on some tasks.[27] Deep learning models exist for analyzing histopathology in a variety of contexts: in bone marrow of humans[15,26] and animals,[11] and other organs, in both humans[6,21–23,28–35] and animals.[36]

The benefits of deep learning for analyzing histopathology WSIs suggested that deep learning would be a promising approach for analyzing bone marrow cellularity in rat histologic WSIs. We considered models used to analyze pathology modalities other than histology, for example bone marrow smear[13] or blood smear[37] and other microscopy modalities.[16,29,34,38] While these models achieved promising performance, they were not amenable to histology WSIs with small hematopoietic cells because the models relied on clear depiction of cell boundaries that was unlike the depiction of small hematopoietic cells in H&E histology.

Deep learning methods use neural networks with many layers to learn image representations iteratively by adjusting network weights to minimize differences between predictions and GT. Deep learning models are applicable to object detection, segmentation, and classification. Two types of segmentation are relevant to our work: semantic segmentation, which partitions scenes into classes by assigning class labels to individual pixels, and instance segmentation, which identifies and labels individual object instances within a given class.[39] Our method is based on 2 network architectures that are commonly used for segmentation: encoder–decoder,[40] and Mask R-CNN.[41] Encoder–decoder architectures are often used for semantic segmentation. The encoder comprises successive layers of convolution operators, activation functions, and pooling operators to create feature-map representations with progressive abstraction. The decoder creates successive upsampled versions of the last encoder feature-map, often with the same pooling indices that the encoder uses to generate downsampled maps. A softmax classification layer produces a pixel-level classification from the last decoder feature map.[40] Mask R-CNN is one of the better performing instance segmentation architectures.[42] The architecture comprises a network that proposes ROIs, a network that extracts features within the ROI bounding boxes and classifies the ROI, a module that refines region bounding box location via regression, and a fully convolutional network that generates a binary mask for each ROI.[41]

[1] We use the term small hematopoietic cells to refer to non-megakaryotic nucleated hematopoietic cells.

[2] The variation in the tinctorial appearance may or may not be related to cell depletion.

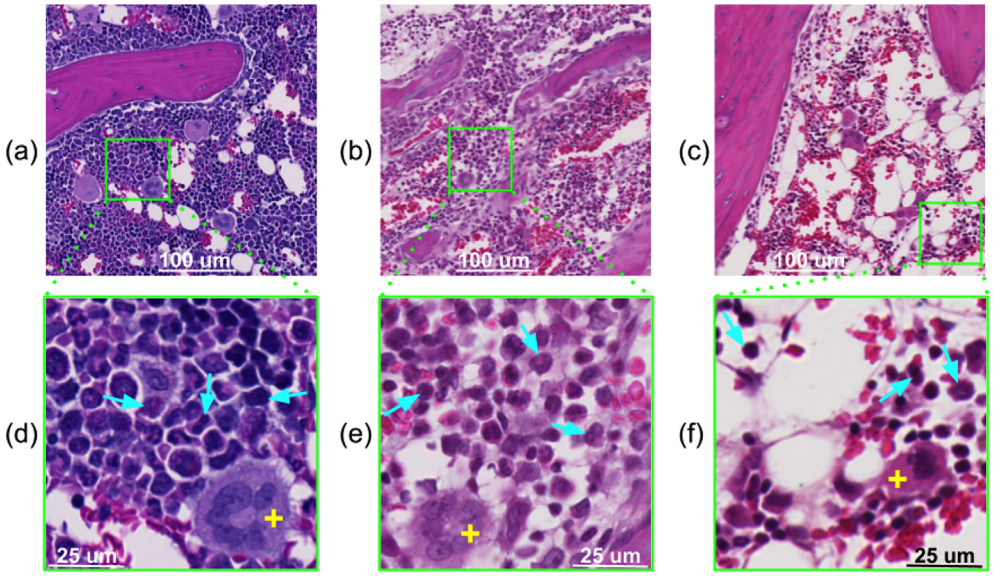

Fig. 2. Variation in the appearance of H&E bone marrow tissue across increasing depletion scores: (a, d) normal (no depletion), (b, e) moderate, and (c, f) severe. MKCs are designated with a yellow '+' sign, and example SHCs with a cyan arrow. As we examine panels (a) through (c) and (d) through (f), the tissue samples appear less-densely packed, consistent with the depletion severity. The SHC examples include cells with a variety of morphologies consistent with the expected diversity of cell lineages and developmental stages comprising the hematopoietic tissue. The tinctorial properties of the MKC cytoplasm vary considerably across the examples shown, with (a, d) exhibiting the most basophilic staining and (c, f) the most eosinophilic staining.

Materials and methods

Framework

We developed a method based on publicly available pre-trained deep learning architectures and updated the model weights using GT data that we collected for target objects used to determine cellularity in rat bone marrow WSIs. Our method includes 4 steps. First, our method identifies the tissue area in the WSI using a *k*-means classifier. Second, the method segments bone marrow regions at 5× magnification, and excludes all non-marrow regions from further analysis. For segmenting bone marrow regions, the method uses a shallow semantic segmentation model.[43] Third, our method segments components within the marrow that are key to further quantification and analysis: MKCs, at 20× magnification, and SHCs, at 40× magnification. For segmenting MKCs and SHCs, the method uses Mask R-CNN[41] instance segmentation models. The method also segments red blood cells (RBCs), at 40× magnification, using a SegNet[40] semantic segmentation model. The models for segmenting marrow components operate independently. Fourth, the method combines the segmentation results from the preceding step into a labeled result mask (Eq. 1). The method uses MKC and RBC segmentation masks to filter out false-positive SHCs (MKC nuclei and RBCs that were falsely segmented as SHCs). From the result mask, the method calculates cell quantities and toxicity endpoints such as MKC and SHC cell densities (Fig. 3). Typical processing time per WSI of size 95,000 × 60,000 pixels is about 12 min on a high-performance computing cluster with a NVIDIA V100 GPU.

$$Result\ mask = Marrow \wedge (MKCs \vee (SHCs \wedge \neg RBCs \wedge \neg MKC)) \quad (1)$$

Ground-truth (GT) collection for training models

For the GT collection, we used male and female (31 Sprague Dawley (Crl:CD(SD)) and 2 Wistar Han (Crl: WI)) rat bone marrow sternum slides from the Genentech study archive, chosen to represent a range of levels of cell depletion. The slides were generated for safety studies between the years 2011 and 2019, processed in 1 of 2 contract research organizations (CROs) using the same protocol, had a thickness of 4–6 μm, and were scanned with a Hamamatsu NanoZoomer-XR or Hamamatsu NanoZoomer-S360 (Hamamatsu Photonics, Hamamatsu City, Japan) at 40× magnification (0.23 μm/pixel). A team of pathologists, computational scientists, and technicians collected GT using annotation tools available in Halo® Image Analysis Platform version v3.2 (Indica Labs). A computational scientist and a technician annotated, on 1 sternebra section per WSI, a boundary around the sternebra, marrow, cartilage, and admixed adipose[3] tissue. Three pathologists and 2 technicians marked all MKC boundaries within the marrow region in 1 sternebra section per WSI. Four pathologists annotated boundaries around small hematopoietic cells in 2 predefined rectangular regions per WSI. Annotating small hematopoietic cells was difficult because of the small size of the cells and the lack of clear boundaries around the cells within the WSIs, especially when cells in the tissue were overlapping or densely positioned. Pathologists reviewed and discussed cases where MKC and SHC boundaries or cell type were difficult to discern. A computational scientist marked boundaries around regions with RBCs. Table 1 presents the details of the GT used to train the models.

Model details

Supplementary Materials Table 1 lists the characteristics of the models included in our method. We trained models with fixed hyperparameters and obtained a baseline intersection over union (IoU)[44,45] via cross-validation for each model. For the bone marrow, MKC, and SHC models, we ran tuning experiments via the tuning platform, Tune,[46] which resulted in only negligible improvement over the baseline IoUs. All models were based on publicly available TensorFlow implementations. The bone marrow and RBC models were based on semantic segmentation model suites,[47,48] the MKC and SHC models were based on a publicly available implementation of Mask R-CNN.[49] All models were trained on tiles (see tile sizes for each model in Table 1). Model inference was conducted as depicted in Fig. 3. Fig. 4 depicts segmentation boundaries for marrow, MKCs, and SHCs. We applied augmentation to the tiles used in the marrow, MKC, and SHC models (see details in Supplementary Material Section 1).

All models included in our method perform inference on partially overlapping tiles, and the resulting segmentation masks for the tiles are stitched

[3] In this paper, we refer to regions of both unilocular and /or multilocular adipocytes, white, brown, and /or beige adipose tissue as admixed adipose tissue.

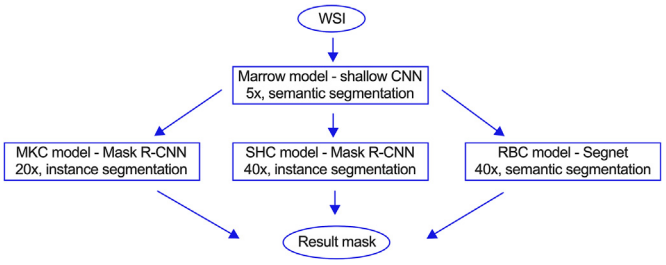

Fig. 3. The model framework used as the basis for the bone marrow analysis pipeline. SHCs that are within the marrow region are filtered by the detected RBCs and then, together with MKCs in the marrow region, are included in the labeled mask.

together to form the result mask for each of the marrow segments. The stitching strategy is dependent on the model type. For the marrow model, every tile mask overrides preceding masks from overlapping tiles. For the MKC, SHC, and RBC models, tile masks are combined with those of preceding tiles by taking the maximum pixel label where pixels overlap. Taking the maximum value allows us to capture all objects that were detected despite potentially different depictions of the same objects in overlapping tiles due to translational variation.[50,51] We observed that for MKC and SHC tile masks, there were cases where objects had duplicate (partial) representations in overlapping areas of neighboring tiles that were cropped by tile edges and were discontinuous in the stitched mask. To alleviate the risk of overcounting, we ignored segments that were very close to tile edges (we assumed these objects would be depicted in overlapping tiles).

Post-processing

Our method applies procedures to address inference errors. For the bone marrow, a procedure removes small regions outside the convex hulls of marrow sections and retains as marrow any holes within the marrow that were likely omitted due to confounding by a similar-looking class of admixed adipose tissue included on many sternebrae bone marrow sections and located outside the marrow cavity (Fig. 5(a)–(c)). For MKCs, a procedure removes small, noisy regions, and false-positive regions with homogeneous colors where nuclei were clearly absent (Fig. 5(d)).

For SHCs, when multiple instances touch, a procedure performs a secondary analysis to determine if instances are actual objects or inference errors. The procedure compares the normalized standard deviation (nSTD) of pixel counts for the touching instances to a threshold of 0.25. A nSTD smaller than the threshold suggests that all instances are actual objects, and the original labels are retained. A nSTD larger than the threshold suggests that the group of instances belongs to one object, and a single instance label is assigned to the touching instances (Fig. 6). We determined the threshold via visual exploration. Large SHC instances are regarded as inference errors where multiple touching cells have a single label. A cell splitting procedure based on the work of Bai et al[52] estimates the correct partitioning of a large SHC instance. Following these procedures for correcting model inferences, our method labels SHC and MKC instances with unique color codes to allow computing end points for the marrow result masks.

Evaluation

A prerequisite to deployment of the bone marrow method as a routine analysis pipeline to aid in safety assessment was a demonstration of good object segmentation accuracy and accurate counts compared to segmentation performed manually by a pathologist and compared to derived counts. We estimated the baseline segmentation accuracy for each model using cross-validation (see Supplementary Material Table 1). Then, we undertook a formal evaluation in which we compared marrow, MKC, and SHC cell segmentation ROIs and cell counts that our method inferred to GT that scientists/pathologists had delineated. We also compared the nuclei counts produced by 2 state-of-the-art deep learning models, Cellpose[23] and Stardist pre-trained with histology weights[53] to the GT. We did not formally evaluate the accuracy of the RBC model alone by comparing segmented RBC regions to RBC GT annotated by pathologists. However, we determined the error rate for the RBC model by comparing the RBC segmentation masks to those of pathologist SHC consensus GT.

Data collection

We selected 1 WSI from each of 10 unique studies that were not used for training the models, all from the Genentech study archive. The slides were Sprague Dawley (Crl:CD(SD)) rat bone sternum slides generated for safety studies between the years 2012 and 2019, processed in 1 of 2 CROs (using the same protocol), had a thickness of 4–6 μm, and were scanned with Hamamatsu NanoZoomer-XR or Hamamatsu NanoZoomer-S360 (Hamamatsu Photonics, Hamamatsu City, Japan) at 40× magnification (0.23 μm/pixel). The selection included WSIs with a range of levels of cell depletion (see slide specification in Supplementary Material). For evaluating the marrow model, a veterinary medical student annotated marrow sections on one of the sternebrae within each of the 10 WSIs using Halo® Image Analysis Platform version v3.2 (Indica Labs). For evaluating the MKC and SHC models, a computational scientist selected 2 ROIs per WSI for SHC annotations (about 65000 μm^2 when combined) and 1 ROI per WSI for MKC annotation (about 8 mm^2, see Supplementary Material Table 6 for MKC and SHC annotation counts). Three board-certified veterinary pathologists from an external CRO, who did not take part in the GT collection for training, annotated all MKCs and SHCs that they identified within all ROIs using QuPath.[54] Fig. 7 shows the annotated ROIs for 1 WSI evaluation. To minimize potential bias related to order of presentation, the order of ROI presentation was randomized for each pathologist. The ROIs were partitioned into 2 groups of 5 (random partition for each pathologist) and were annotated by the pathologist in 2 sessions with a month gap between the groups to minimize pathologist fatigue. We established the pathologist consensus GT by aligning, for a given ROI, the 3 annotation masks generated by the pathologists, and selecting only pixels that were included in at least 2 out of the 3 masks. In order to estimate intra-pathologist variation, after a washout period of at least 6 months (to eliminate potential recollection of preceding annotations), the annotation process was repeated with the same pathologists and ROIs.

Tests and metrics

To compare the bone marrow segmentations obtained using our method with segmentations determined by scientists, we computed the

Table 1
GT used for training models.

Tissue type	Tile size	WSI/Tiles per depletion severity						Unique studies	Total count
		Normal	Minimal	Mild	Moderate	Marked	Severe		
Marrow	132	7/439	5/565	5/657	0	3/599	3/595	15	17[a]
MKCs	1024	2/54	2/89	0	1/31	2/31	1/36	8	1597[b]
SHCs	128	2/122	2/115	0	0	2/248	1/60	7	7114[b]
RBCs	128	5/435	3/553	1/13	0	0	0	8	1111[c]

[a] Sternebra sections.
[b] Cells.
[c] ROIs.

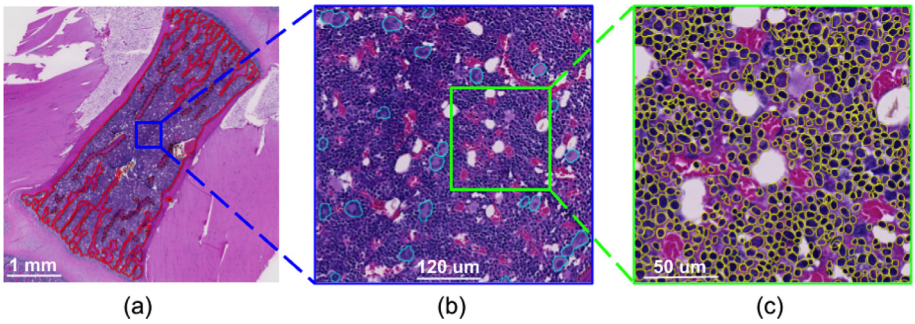

Fig. 4. Segmentation results of the bone marrow analysis pipeline for normal tissue, depicted with contours around the segmented objects. (a) Marrow within a sternebra section (red). The dashed red lines depict bone, which is excluded from the marrow. (b) A marrow patch with segmented MKCs (cyan). (c) A marrow patch with segmented SHC (yellow). In this example, there was no pixel overlap between the SHC and RBC masks. Thus, panel c is the same before and after application of the RBC mask.

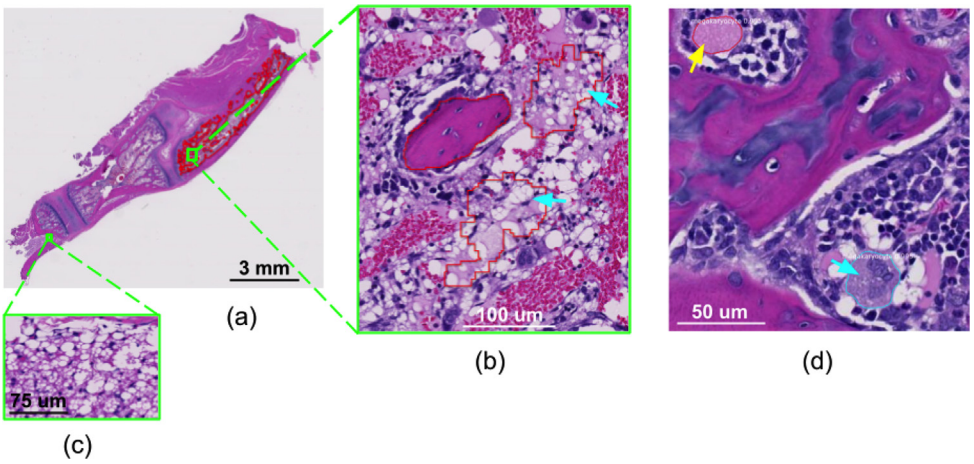

Fig. 5. Example of segmentation error for the marrow and MKC models that were addressed with post-processing. (a) A WSI with marked cell depletion. (b) Regions within the marrow that the model confused with adipose tissue (cyan arrows). (c) Adipose tissue. (d) MKC segmentation in an ROI of a WSI with no cell depletion (yellow and cyan arrows). The region pointed to by the yellow arrow represents a segmentation error because it includes a nuclear material that did not meet the morphologic criteria for designation as an MKC and is in contrast to the nucleated megakaryocyte segmentation pointed to by the cyan arrow.

Dice score,[44,45] which is based on the coordinate overlap of 2 ROIs (twice the number of pixels in the intersection between the ROIs divided by the total number of pixels in both ROIs). To compare the method's cell segmentations to pathologist segmentations, and to establish inter- and intra-pathologist variability, we computed the Dice and the object Dice[55] scores, the second of which estimated per-object segmentation overlap.[55] In order to evaluate the impact of discrepancies in object segmentation on derived total cell counts, we compared for MKCs and SHCs cell counts derived from our method to cell counts derived from pathologist annotations with the Bland–Altman test.[56] For SHCs, we also compared counts derived from Cellpose and Stardist nuclei segmentations to cell counts derived from pathologist annotations. The comparison of SHC cell counts derived from Cellpose and Stardist nuclei segmentations were based on 2 configurations: 1 with the filtering of RBC and MKC false-positives and 1 without. The Bland–Altman test was suitable for cell count comparisons because it established the agreement interval between counts derived from 2 imperfect segmentation methods, our automated method and pathologist delineations. To determine the RBC error rate we calculated, for each evaluation ROI, the percentage of RBC pixels that overlapped with SHC pathologist consensus GT.

Results

Comparison of segmentations from our method to those of scientists/pathologists

The mean Dice score for our method's marrow segmentation vs scientist bone marrow segmentations was 0.9 ± 0.1. Box plots of mean Dice and mean object Dice scores for our method's cell segmentations vs pathologist cell segmentations, and for pathologist pairs (estimate of the inter- and intra-pathologist variation for cell segmentations) are shown in Fig. 8 (Dice and object Dice scores for individual WSI appear in Supplementary Material tables 2–5). The median Dice and median object Dice scores determined for MKCs using our method vs pathologist consensus and for inter- and intra-pathologist variation were comparable with overlapping first-third quartile ranges. For SHCs, the median scores were close, with first-third quartile ranges partially overlapping inter-pathologist variation. The error rate for RBC segments was 9%. We observed that SHC Dice scores for our method when applied with RBC false-positive filtering were the same as Dice scores obtained when our method was applied without RBC filtering, the mean Dice score was 0.70 in both cases (see Supplementary Material Table 4).

Comparison of cell counts derived from cell segmentation methods and cell segmentations of pathologists

The Bland–Altman plots (Fig. 9) show the difference between cell counts derived from automated segmentation and from pathologist segmentation consensus for the evaluation ROIs. The average consensus GT cell count per evaluated slide (1 ROI for MKC, 2 ROIs for SHC) for MKCs and SHCs were 28 and 431, respectively. The 95% limits of agreement designate the range of values that encompasses the differences for most ROIs. The mean difference between our method and pathologist consensus for the MKC counts was 6.2 with 95% limits of agreement ranging from -14 to 27 (Fig. 9(a)), and a percent difference of 22% relative to the average cell count.

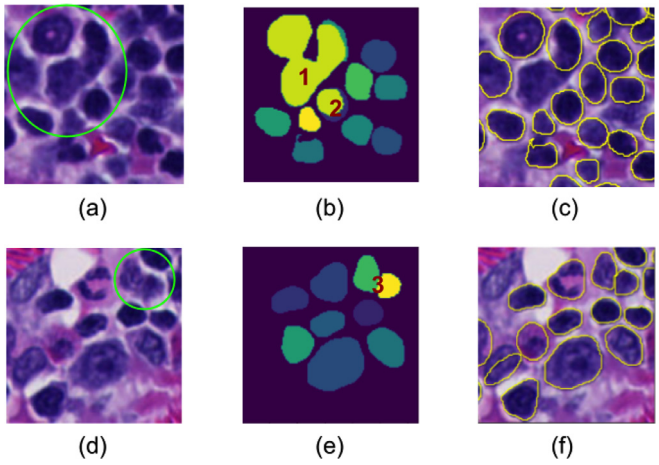

Fig. 6. Typical tiles with SHC instance segmentation errors partially resolved by post processing. (a, d) Raw tiles with visible cells. (b, e) Segmentation masks for the tiles, depicting instances in different colors (note that the colors of some instances may look similar despite having different color values). (c, f) Result cell boundaries after post-processing superimposed on raw tiles. Note that boundaries of cells along the tile perimeter, which are not depicted in the masks, are included in masks of neighboring overlapping tiles. Missing cells are a result of either segmentation errors or errors in the cell splitting procedure. (a–c) Segmentation errors addressed by secondary nSTD analysis or cell splitting. (b1) Shows 3 cells segmented as 2 instances. (b2) Shows 1 cell segmented as 2 instances. (d–f) Touching instances that were correctly determined to be 2 cells by the secondary nSTD analysis. (e3) Shows 2 cells segmented as 2 instances.

The results shown in Fig. 9(b–d) for SHC counts reflect cell counts derived from segmentation masks after filtering false-positives (see example Supplementary Material Fig. 1). The mean difference between our method and pathologist consensus for the SHC counts was 62.9 with 95% limits of agreement ranging from −15 to 140 (Fig. 9(b)), and a percent difference of 15% relative to the average cell count. The percent cell count differences for MKCs and SHCs indicate that our method to pathologist differences were more prominent for MKCs than for SHCs. In Fig. 9(c), the mean difference between the Cellpose and pathologist consensus SHC counts was 82.7 with 95% limits of agreements ranging from −170 to 330 (Fig. 9(c)), and a percent difference of 19% relative to the average cell count. Finally, the mean difference between the Stardist and pathologist consensus SHC counts was 120.4 with 95% limits of agreement ranging from −45 to 290 (Fig. 9(d)), and a percent difference of 28% relative to the average cell count. Based on these results, our method produced counts that were closer to pathologist counts than Cellpose and Stardist (Fig. 9(b)–(d)), with a smaller 95% limits of agreement range.

False-positive filtering for Cellpose resulted in a cell count that was closer to that of the pathologist consensus GT than cell count without false-positive filtering. For Stardist, false-positive filtering resulted in a cell count that was slightly farther apart from the pathologist consensus GT. However, SHC cell counts for our method were closest to pathologist consensus GT compared to the other 2 methods (with or without false-positive filtering, see Supplementary Materials Tables 7–9). A visual example of SHC segmentations in 1 evaluation ROI for the consensus GT and all 3 SHC segmentation models is shown in Supplementary Material Fig. 1.

Discussion

The scientific literature includes a few reports of automated methods for analyzing WSIs of histologic preparations of bone marrow samples from animal models,[7,11,12,57] all of which could potentially enhance the work of toxicologic pathologists. However, these methods were not suitable for routine high-throughput cellularity analysis of bone marrow histology in new studies, either because the methods were not targeted at assessing cellularity,[12] because the foundational software was no longer available,[7] the methods were designed for animal models other than rat,[11,12] or because the methods were evaluated on only 1 study and therefore were not demonstrated to be generalizable.[11] We considered adapting available models for analyzing human bone marrow to rat data, which was similar to human bone marrow,[58] but preferred to tailor our models to rat data to increase the likelihood of achieving highly accurate results. We embarked upon the task of collecting GT data for model training and evaluation and trained multiple publicly available semantic and instance segmentation models to identify key tissue components while removing irrelevant components.

We observed that segmenting MKCs was easier than segmenting SHCs, both for pathologists and for our method, as reflected in the higher Dice and object Dice scores for MKCs. Segmenting SHCs was most difficult when the cells were densely arranged and overlapped. Part of the challenge stemmed from the difficulty discerning cytoplasm borders, which did not always stain well in H&E slides, and distinguishing between overlapping cells, together making cell boundaries difficult to visualize. Unlike other projects that analyzed tissue with small cells, where researchers resorted to nuclei segmentation,[7,59] we performed our analysis with whole cells to capture

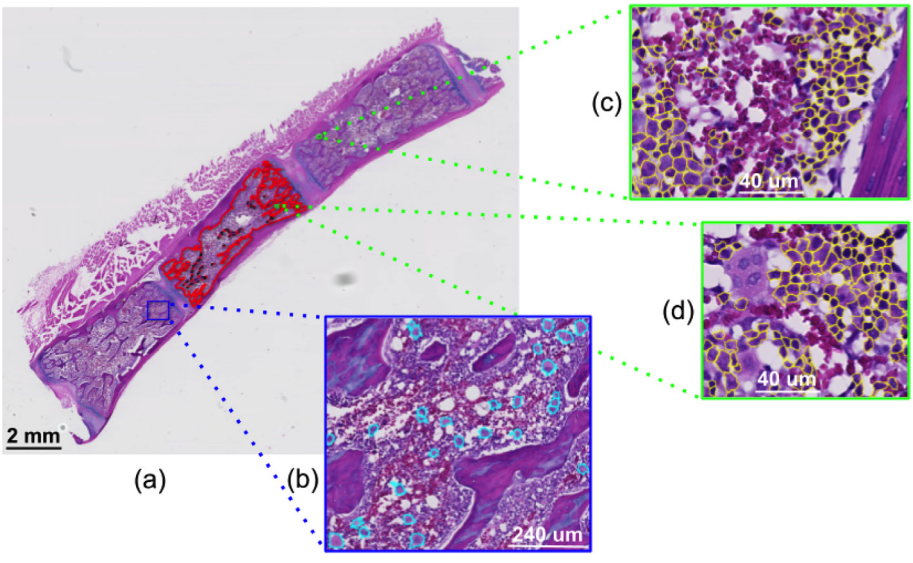

Fig. 7. Manually annotated ROIs in a WSI with marked depletion. (a) A view of the entire WSI with marrow annotations around a sternebra section (red), and demarcation of MKC (blue) and SHC (green) ROIs. (b–d) Zoomed in views of MKC and 2 SHC ROIs, respectively. MKCs are shown in cyan (b), SHCs in yellow (c, d).

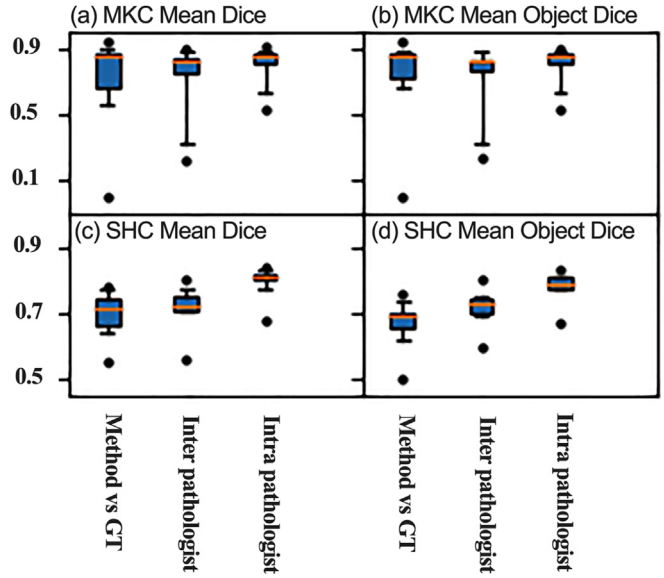

Fig 8. Box plots for mean Dice and mean object Dice scores for cell segmentation. (a–b) Scores for MKC segmentation. (c–d) Scores for SHC segmentation. In all plots, the whiskers indicate the 5 and 95 percentiles, and extreme outliers are plotted as circles. The orange horizontal lines depict medians. Note that for MKC, the lowest Dice and object Dice values were zero because the 2 WSIs with severe cell depletion had few MKCs, and there was no full agreement with respect to these MKCs. The term GT in the plot tick labels refers to consensus GT. The data from which the plot was generated appear in Supplementary Material Tables 2 through Table 5.

the highly variable nuclear morphologies, for example, band, u-shaped, and lobulated nuclear forms of cells in the granulocytic lineages. We anticipated that the bone marrow nuclear morphologies would be challenging for nuclear segmentation algorithms optimized to identify oval shaped nuclei with more homogeneous size and shape characteristics. We combined overlapping tiles with a maximum operation to consider all information that models inferred in overlapping areas, some of which was not consistent due to translational variation.[50,51] Our post-processing analysis based on nSTD helped us identify and merge multiple instances for single objects. We found that SHC cell counts derived from Mask R-CNN inference were closer to those derived from GT compared to counts derived from Cellpose and Stardist inferences. For MKCs and SHCs, small cell count differences between our method and pathologist scores held across different pathological contexts selected for the evaluation slides. We observed that in the current version of our method, filtering of RBC false-positives did not improve the accuracy of SHC segmentation compared to pathologist consensus GT. This result is not surprising given that the accuracy of the RBC model was limited. However, we anticipated that future improvements to the RBC model would enhance the performance of our method and therefore we retained the RBC model in our framework.

Although deployment of publicly available segmentation models was much easier than collecting GT for rat-bone marrow WSIs, our goal to achieve accuracy close to that of pathologists justified the GT collection effort. The GT specific to rat bone marrow yielded segmentation and cell count accuracies that were close to those of pathologists, and a robust and consistent pipeline for high-throughput screening of rat bone marrow WSIs for cell depletion.

Conclusion

The strong performance of the bone marrow analysis pipeline supports its incorporation into routine use as an aid for hematotoxicity assessment

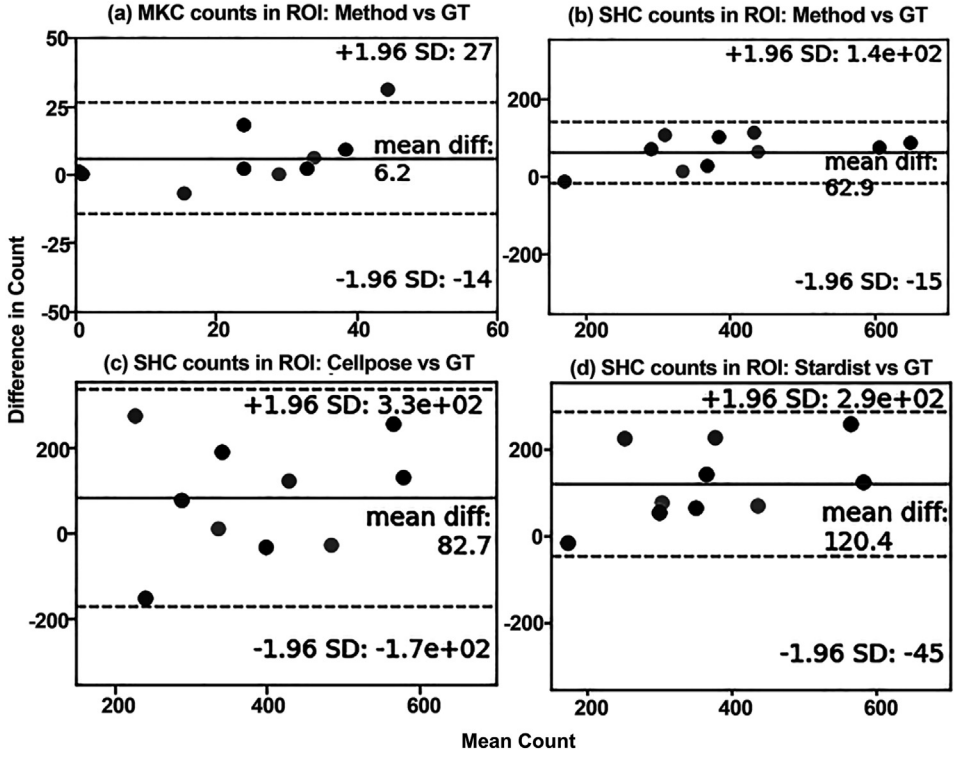

Fig. 9. Bland–Altman plots for cell count comparisons in ROIs from 10 evaluation slides, chosen from 10 different studies to represent a range of levels of cell depletion. (a) Our method's MKC counts vs pathologist MKC consensus counts. (b) Our method's SHC counts vs pathologist SHC consensus counts. (c) Cellpose SHC counts vs pathologist SHC consensus counts. (d) Stardist SHC counts vs pathologist SHC consensus counts. Positive differences indicate that the GT count was higher than the count of the compared method, negative differences indicate that the GT count was lower than the count of the compared method. The data from which the plot was generated appear in Supplementary Material Table 6.

by pathologists. To build diagnostic confidence in the method for use by pathologists, we will derive scores from pipeline endpoints and compare these to pathologist bone marrow scores in studies. We will also compare differences between study groups derived from our methods to differences between study groups derived from pathologist scores. Moreover, we will define a standard procedure for assessing pipeline analysis results to assure quality of pipeline performance on every sample. An important next step for our work would be to test the bone marrow analysis pipeline on rat WSI generated and scanned at other sites (beyond the 2 CROs that processed the slides used in our work), and to address compatibility issues that arise by applying domain adaptation techniques to our data. This effort is underway.

We envision that utilizing our method may enable meta-analysis of rat bone marrow characteristics from future and historical WSIs and may generate new biological insights from cross-study comparisons. We anticipate continuing to improve segmentation accuracy by generating additional GT annotations and retraining the method's models. We hope to reduce the RBC error rate so that we can improve the accuracy of our method by filtering out false-positive SHC. Potential extensions of our work include adapting the models to H&E bone marrow from additional bone tissues routinely collected in safety studies, such as the femur, as well as applying the approach to other preclinical species. Additional future goals include training models that can identify developmental cell states and distinguish between the hematopoietic cell's lineages. We are looking to facilitate manual collection of segmentation GT by pairing bone marrow histology WSIs with WSIs depicting hematopoietic cells identified with immunohistochemical markers, and to leverage semi-supervised learning methods for automatic identification of hematopoietic cell annotations using existing GT.

Funding source

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of Competing Interest

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: all authors were employees of Roche/Genentech at the time the work described in the article was performed. Some of the authors own Roche stock.

Acknowledgments

We thank Trung Nguyen and Serena Ngo for scanning slides and training ground-truth annotation, Debra Tokarz, Thomas Steinbach, Kathleen Funk for ground-truth annotations for algorithm evaluation, Philip Shen for prototyping the RBC model in Halo, and Cleopatra Kozlowski, Andries Zijlstra, Fangyao Hu for collegial discussions and feedback.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2023.100333.

References

1. Biddle KE. Opinion on the optimal histologic evaluation of the bone marrow in nonclinical toxicity studies. Toxicol Pathol 2021;50:266-273. https://doi.org/10.1177/01926233211061712.
2. Boes KM, Durham AC. Bone marrow, blood cells, and the lymphoid/lymphatic system. Pathologic Basis of Veterinary Disease; 2017. p. 724-804.e2.
3. Ramaiah L, Bounous DI, Elmore SA. Hematopoietic system. Haschek and Rousseaux's Handbook of Toxicologic Pathology; 2013. p. 1863-1933.
4. Reagan WJ, Irizarry-Rovira A, Poitout-Belissent F, et al. Best practices for evaluation of bone marrow in nonclinical toxicity studies. Vet Clin Pathol 2011;40:119-134. https://doi.org/10.1111/j.1939-165X.2011.00323.x.
5. Valli VEO, Kiupel M, Bienzle D, Wood RD. Hematopoietic system. Jubb, Kennedy & Palmer's Pathology of Domestic Animals. 2016. p. 102-268. 3. e1.
6. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology image analysis using segmentation deep learning algorithms. Am J Pathol 2019;189:1686-1698. https://doi.org/10.1016/j.ajpath.2019.05.007.
7. Kozlowski C, Brumm J, Cain G. *An Automated Image Analysis Method to Quantify Veterinary Bone Marrow Cellularity on H&E Sections Toxicologic Pathology*, 46. 2018:324-335. https://doi.org/10.1177/0192623318766457.
8. Holzinger A, Malle B, Kieseberg P, et al. Machine learning and knowledge extraction in digital pathology needs an integrative approach. Towards Integrative Machine Learning and Knowledge Extraction; 2017. p. 13-50.
9. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform 2018;9. https://doi.org/10.4103/jpi.jpi_53_18.
10. Gallas BD, Chan H-P, D'Orsi CJ, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. Acad Radiol 2012;19:463-477. https://doi.org/10.1016/j.acra.2011.12.016.
11. Smith MA, Westerling-Bui T, Wilcox A, Schwartz J. Screening for bone marrow cellularity changes in cynomolgus macaques in toxicology safety studies using artificial intelligence models. Toxicol Pathol 2021;49:905-911. https://doi.org/10.1177/0192623320981560.
12. Tratwal J, Bekri D, Boussema C, et al. *MarrowQuant Across Aging and Aplasia: A Digital Pathology Workflow for Quantification of Bone Marrow Compartments in Histological Sections Frontiers in Endocrinology*, 11. 2020. https://doi.org/10.3389/fendo.2020.00480.
13. Wang C-W, Huang S-C, Lee Y-C, Shen Y-J, Meng S-I, Gaol JL. Deep learning for bone marrow cell detection and classification on whole-slide images. Med Image Anal 2022;75. https://doi.org/10.1016/j.media.2021.102270.
14. Hu B, Tang Y, Chang EIC, Fan Y, Lai M, Xu Y. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. IEEE J Biomed Health Inform 2019;23:1316-1328. https://doi.org/10.1109/jbhi.2018.2852639.
15. Song T-H, Sanchez V, Ei Daly H, Rajpoot NM. Simultaneous cell detection and classification in bone marrow histology images. IEEE J Biomed Health Inform 2019;23:1469-1476. https://doi.org/10.1109/jbhi.2018.2878945.
16. Akram SU, Kannala J, Eklund L, Heikkila J. Cell proposal network for microscopy image analysis. 2016 IEEE International Conference on Image Processing (ICIP); 2016. p. 3199-3203.
17. Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. IEEE Rev Biomed Eng 2014;7:97-114. https://doi.org/10.1109/rbme.2013.2295804.
18. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Computat Struct Biotechnol J 2018;16:34-42. https://doi.org/10.1016/j.csbj.2018.01.001.
19. Zuraw A, Staup M, Klopfleisch R, et al. Developing a qualification and verification strategy for digital tissue image analysis in toxicological pathology. Toxicol Pathol 2020;49:773-783. https://doi.org/10.1177/0192623320980310.
20. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal 2019;58. https://doi.org/10.1016/j.media.2019.101544.
21. Mormont R, Geurts P, Maree R. Multi-task pre-training of deep neural networks for digital pathology. IEEE J Biomed Health Inform 2021;25:412-421. https://doi.org/10.1109/jbhi.2020.2992878.
22. Raza SEA, Cheung L, Shaban M, et al. Micro-Net: a unified model for segmentation of various objects in microscopy images. Med Image Anal 2019;52:160-173. https://doi.org/10.1016/j.media.2018.12.003.
23. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nat Methods 2020;18:100-106. https://doi.org/10.1038/s41592-020-01018-x.
24. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng 2009;2:147-171. https://doi.org/10.1109/rbme.2009.2034865.
25. Nielsen FS, Pedersen MJ, Olsen MV, Larsen MS, Røge R, Jørgensen AS. Automatic bone marrow cellularity estimation in H&E stained whole slide images. Cytometry Part A 2019;95:1066-1074. https://doi.org/10.1002/cyto.a.23885.
26. Song T-H, Sanchez V, ElDaly H, Rajpoot NM. Hybrid deep autoencoder with Curvature Gaussian for detection of various types of cells in bone marrow trephine biopsy images. 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017); 2017. p. 1040-1043.
27. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. 2015 IEEE International Conference on Computer Vision (ICCV); 2015. p. 1026-1034.
28. Chen H, Qi X, Yu L, Dou Q, Qin J, Heng P-A. DCAN: deep contour-aware networks for object instance segmentation from histology images. Med Image Anal 2017;36:135-146. https://doi.org/10.1016/j.media.2016.11.004.
29. Falk T, Mai D, Bensch R, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods 2018;16:67-70. https://doi.org/10.1038/s41592-018-0261-2.
30. Hägele M, Seegerer P, Lapuschkin S, et al. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Scient Rep 2020;10. https://doi.org/10.1038/s41598-020-62724-2.
31. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 2016;7. https://doi.org/10.4103/2153-3539.186902.
32. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B. A generalized deep learning framework for whole-slide image segmentation and analysis. Scient Rep 2021;11. https://doi.org/10.1038/s41598-021-90444-8.
33. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170-175. https://doi.org/10.1016/j.media.2016.06.037.

34. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation medical image computing and co mputer-assisted intervention–MICCAI 2015. 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18; 2015. p. 234-241. Conference: Location: Springer.
35. Zhang Z, Chen P, McGough M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning Nature. Mach Intel 2019;1:236-245. https://doi.org/10.1038/s42256-019-0052-1.
36. Hoefling H, Sing T, Hossain I, et al. HistoNet: a deep learning-based model of normal histology. Toxicol Pathol 2021;49:784-797. https://doi.org/10.1177/0192623321993425.
37. Kassim YM, Palaniappan K, Yang F, et al. Clustering-based dual deep learning architecture for detecting red blood cells in malaria diagnostic smears. IEEE J Biomed Health Inform 2021;25:1735-1746. https://doi.org/10.1109/jbhi.2020.3034863.
38. Arbelle A, Raviv TR. Microscopy cell segmentation via adversarial neural networks. IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018 April 7th; Conference: Location: IEEE; 2018. p. 645-648.
39. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:170406857; 2017.
40. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intel 2017;39:2481-2495. https://doi.org/10.1109/tpami.2016.2644615.
41. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 2980-2988.
42. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. *Image Segmentation Using Deep Learning: A Survey*. 2020.p. arXiv:2001.05566.
43. Lei F, Liu X, Dai Q, Ling BW-K. Shallow convolutional neural network for image classification. SN Appl Sci 2019;2. https://doi.org/10.1007/s42452-019-1903-4.
44. Eelbode T, Bertels J, Berman M, et al. Optimization for medical image segmentation: theory and practice when evaluating with Dice Score or Jaccard Index. IEEE Trans Med Imaging 2020;39:3679-3690. https://doi.org/10.1109/tmi.2020.3002417.
45. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15. https://doi.org/10.1186/s12880-015-0068-x.
46. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: a research platform for distributed model selection and training. arXiv preprint arXiv:180705118; 2018.
47. Lu Y. Amazing Semantic Segmentation. [cited 2022 Sep. 26, 2022] Available from: https://github.com/luyanger1799/Amazing-Semantic-Segmentation 2020.
48. Seif G. Semantic Segmentation Suite in TensorFlow. [cited 2022 26 Sep 2022] Available from: https://github.com/GeorgeSeif/Semantic-Segmentation-Suite 2019.
49. Abdulla W. Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow. [cited] Available from: https://github.com/matterport/Mask_RCNN 2017.
50. Huang B, Reichman D, Collins LM, Bradbury K, Malof JM. Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations. arXiv preprint arXiv:180512219; 2018.
51. Reina GA, Panchumarthy R, Thakur SP, Bastidas A, Bakas S. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. Front Neurosci 2020;14. https://doi.org/10.3389/fnins.2020.00065.
52. Bai X, Sun C, Zhou F. Splitting touching cells based on concave points and ellipse fitting. Pattern Recognit 2009;42:2434-2446. https://doi.org/10.1016/j.patcog.2009.04.003.
53. Schmidt U, Weigert M, Broaddus C, Myers G. Cell detection with star-convex polygons. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018; 2018. p. 265-273.
54. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. Scient Rep 2017;7. https://doi.org/10.1038/s41598-017-17204-5.
55. Sirinukunwattana K, Pluim JPW, Chen H, et al. Gland segmentation in colon histology images: the glas challenge contest. Med Image Anal 2017;35:489-502. https://doi.org/10.1016/j.media.2016.08.008.
56. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 2016;8:135-160. https://doi.org/10.1177/096228029900800204.
57. Kozlowski C, Fullerton A, Cain G, Katavolos P, Bravo J, Tarrant JM. Proof of concept for an automated image analysis method to quantify rat bone marrow hematopoietic lineages on H&E sections. Toxicol Pathol 2018;46:336-347. https://doi.org/10.1177/0192623318766458.
58. Ward JM, Cherian S, Linden MA. Hematopoietic and lymphoid tissues. Comparative Anatomy and Histology; 2018. p. 365-401.
59. Song TH, Sanchez V, EI H, Rajpoot NM. Dual-channel active contour model for megakaryocytic cell segmentation in bone marrow trephine histology images. IEEE Trans Biomed Eng 2017;64:2913-2923. https://doi.org/10.1109/TBME.2017.2690863.