MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

ORIGINAL ARTICLE

Optimizing Storage and Computational Efficiency: An Efficient Algorithm for Whole Slide Image Size Reduction

Shahriar Faghani, MD; D. Chamil Codipilly, MD; Mana Moassefi, MD; Prasad G. Iyer, MD; and Bradley J. Erickson, MD, PhD

Abstract

Objective: To efficiently store, transfer, and analyze whole slide imaging (WSI), we developed an image-processing algorithm to remove the unneeded background in a WSI and assemble tissue-containing parts into smaller WSIs without any change in tissue area image resolution.

Patients and Methods: We used histology slides of nondysplastic Barrett esophagus, low-grade dysplasia, and high-grade dysplasia, which were digitized using Aperio AT2 Scanner from January 1992 to September 2020. The algorithm involved converting color images to grayscale images, binarizing images by assigning zero to the background and 1 to the foreground, filling the holes and dilating the foreground masks, and extracting connected components. Using the coordinates of each component, the vertices of the smallest surrounding bounding box were calculated, and tissue-containing parts were cropped from the original slide. The smallest possible rectangle that encloses all bounding boxes containing tissue was found using the rectangle-packer package.

Results: The algorithm resulted in a mean reduction of 7.11×. The performance of a previously developed deep learning model for the detection of Barrett esophagus dysplasia grade on the size-reduced WSIs was comparable with that on the original WSIs.

Conclusion: Our algorithm for WSI size reduction can assist researchers in storing, transferring, and analyzing WSIs while optimally using them in their workflow.

■ Mayo Clin Proc Digital Health 2023;1(3):419-424

The digital pathology field has grown exponentially, largely driven by advances in imaging hardware and computational power.[1] Digitization of entire histology slides is now being adopted across the world in pathology laboratory workflows.[2] Digital pathology and whole slide imaging (WSI) have revolutionized the way pathology is practiced today and present new opportunities for developing algorithms and software tools to assist pathologists, in both clinical and research settings. Pathologists are able to view and analyze slides remotely and collaborate with other pathologists. In addition, WSIs are widely accepted in education as alternatives to conventional slides because they provide many students with full annotation possibilities, without the problem of serial section variation. Regarding clinical applications, image-processing techniques can also be applied to WSI, providing pathologists with tools to aid in diagnosis.[3] Finally, digital pathology has benefited greatly from the use of conventional machine learning and deep learning algorithms. The most important advantage of applying deep learning in pathology is to reduce errors in diagnosis and classification.[4,5] It is accepted that deep learning—based models in pathology will become part of the new standard of care along with clinical information, biomarkers, and multiomics data.[5,6] Applications of deep learning algorithms in digital pathology have been remarkably successful.[7-11]

WSI, as an important component of digital pathology adoption, involves 2 steps: creating digital images of a glass histopathology or cytopathology slide, followed by viewing a large-sized digital image with a virtual slide viewer.[12,13] However, with the advent of

From the Artificial Intelligence Laboratory, Department of Radiology (S.F., M.M., B.J.E.), and Barrett's Esophagus Unit, Division of Gastroenterology and Hepatology (C.C., P.G.I.), Mayo Clinic, Rochester, MN.

WSI comes a new set of challenges, one of which is the excessive size of the digitized slides. A single WSI can have several gigabytes of data, which can be difficult to store and transfer, especially for pathology departments that have limited storage space and bandwidth. Additionally, large file sizes can affect loading and viewing of slides because specialized software and hardware may be needed to handle the high-resolution images. Furthermore, the size of WSIs can also impair image analysis because certain algorithms may not be able to handle high-resolution images or may not be optimized for large-scale image analysis. This can limit the use of deep learning and machine learning in digital pathology, which otherwise has the potential to greatly improve diagnostic accuracy, efficiency, and education.

Although digital pathology and WSI have many benefits, the huge size of the digital slides presents a number of challenges that must be addressed, such as issues with storage, transfer, load, and image analysis. A considerable part of these sizable slides consists of regions outside the tissue area (so called "white-space" without any clinically relevant information). In this study, we developed an image-processing algorithm that can remove the unneeded background in a WSI. Then, WSI's tissue-containing parts are kept and assembled into a smaller WSI with no change in tissue area image resolution. To assess the output quality, we ran a previously developed deep learning model and compared the results with original WSIs. This strategy for WSI size reduction can assist researchers in storing, transferring, and analyzing WSIs while optimally using them in their workflow.

METHOD

We obtained histology slides of nondysplastic Barrett esophagus, low-grade dysplasia, and high-grade dysplasia (hematoxylin and eosin—stained) from our institution's clinical pathology archive from 1992 to 2020. These slides were digitized (Aperio AT2 Scanner; Leica Biosystems).

Images in the tagged image format were loaded as NumPy arrays using the Tifffile package.[14] Using the morphology and color module from the scikit-image package (v 0.19.3) on Python 3.9, for each slide, first, we converted the color images into the

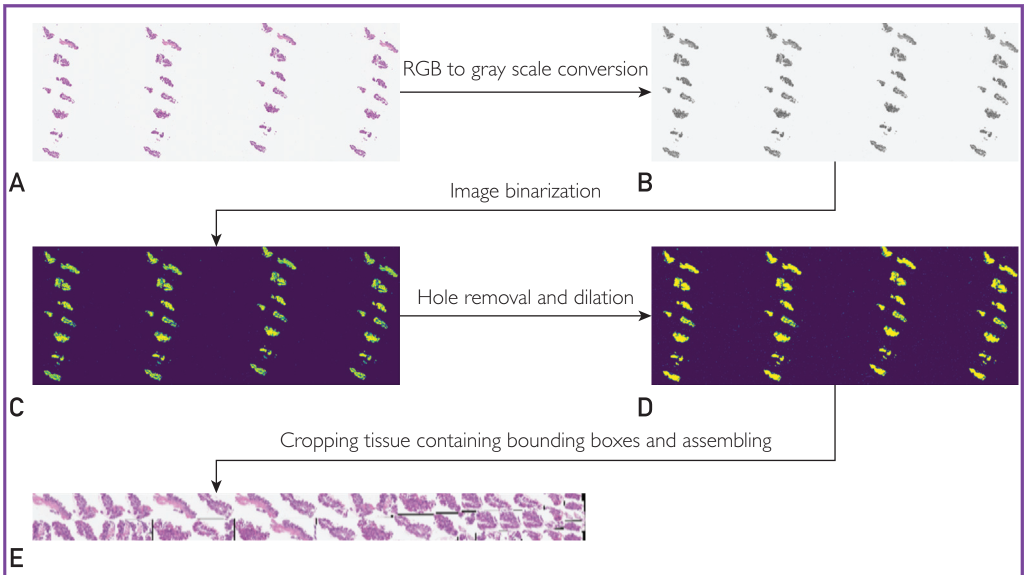
RGB to gray scale conversion
A
B
Image binarization
Hole removal and dilation
C
D
Cropping tissue containing bounding boxes and assembling
E

FIGURE 1. Summary of image-processing steps used in size reduction algorithm. (A) Original red-green-blue (RGB) whole slide image (WSI) is converted to (B) grayscale WSI. Grayscale WSI then is transformed to (C) binary image. (D) Foreground masks are expanded, and holes are filled. (E) Using the foreground masks the tissue-containing parts are cropped and packed into the smallest possible rectangle.

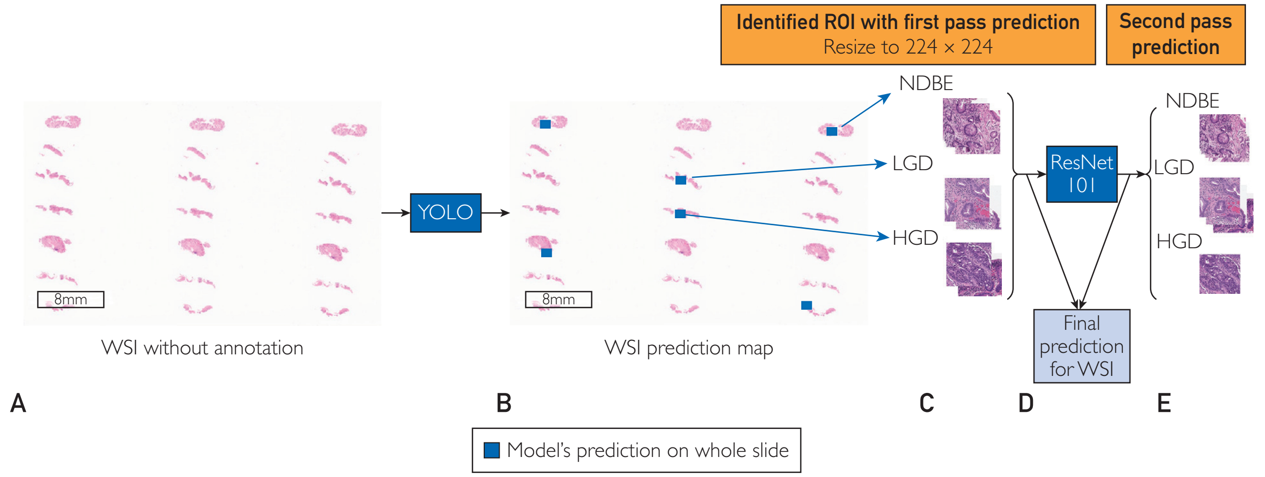

FIGURE 2. Summary of inference pipeline of Barrett esophagus dysplasia deep learning model (BEDDLM). (A) Removal of pathology annotations from test set slides with first-pass object detection model assessment. (B) Regions of interest (ROI) as identified by the object detection model. (C) Resizing of ROI to 224 × 224 pixels for feeding into object detection mode. (D) Second-pass prediction with the classifier model. (E) Final dysplasia grade prediction by ensemble model. HGD, high-grade dysplasia; LGD, low-grade dysplasia; NDBE, nondysplastic Barrett esophagus; WSI, Whole slide imaging; YOLO, you only look once.

grayscale images; then, applying a threshold, we binarized the images by assigning zero to the background and 1 to the tissue-containing part (foreground).[15] We filled the holes and dilated the foreground masks, and using the measure module from scikit-image package (v 0.19.3), connected components were extracted. Using the coordinates of each component, the vertices of the smallest surrounding bounding box were calculated. Furthermore, using the bounding box coordinates, tissue-containing parts were cropped from the original slide (Figure 1A-D).

Finding a rectangle with the smallest area that includes all smaller bounding boxes can be considered an optimization problem with 2 constraints: the total area of the bounding boxes and the shape of each bounding box. We looked at this problem as a variation of the 2-dimensional Knapsack problem, an optimization problem in which there is a limited weight capacity and multiple items with different weights and values, with the goal of finding the combination of items that maximizes the total value while keeping the total weight under the limit.[16]

To find the smallest rectangle that contains all bounding boxes, we used the rectangle-packer package, which provides a solution for solving the 2-dimensional Knapsack problem.[17]

After finding the smallest possible rectangle (new slide) that encloses all bounding boxes that contain tissue, a NumPy array with the size of new slides was created and filled with cropped bounding boxes containing tissue) (Figure 1E).

To preserve the spatial positioning of the bounding boxes, we store their coordinates in a JavaScript Object Notation file. This allows for easy recovery of the bounding box information at a later time if needed.

To ensure compatibility with WSI viewers and achieve a pyramidal resolution, we used ImageMagick (v 7.1.1) to convert the final NumPy array to tagged image format. The conversion was performed with a tile size of 256 × 256.

The mean and SD of heights and widths of WSIs before and after applying the size reduction algorithm was reported. Moreover, the mean and SD of the compression factor were calculated.

To assess the effect of the size reduction algorithm on the performance of downstream machine learning tasks, we fed the size-reduced WSIs to the previously developed model for the detection of Barrett esophagus

TABLE. Comparative Analysis of Barrett Esophagus Dysplasia Detection Machine Learning Model Performance with and without Size Reduction Algorithm

Class	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	F1 score
BEDDLM applied on original WSIs					
NDBE (n=18)	94.4 (72.7-99.8)	96.2 (86.7-99.5)	99.2 (97.2-99.8)	75.3 (31.2-95.3)	.919
Low-grade dysplasia (n=33)	81.8. (64.5-93)	97.3 (85.8-99.9)	77 (32.5-95.9)	97.9 (95.9-99)	.885
High-grade dysplasia (n=19)	94.7 (73.9-99.8)	90.2 (78.5-96.7)	33.7 (18-54.9)	99.6 (97.9-99.9)	.857
BEDDLM applied on size-reduced WSIs					
NDBE (n=18)	94.4 (72.7-99.8)	96.2 (86.7-99.5)	99.2 (97.2-99.8)	75.3 (31.2-95.3)	.919
Low-grade dysplasia (n=33)	81.8. (64.5-93)	97.3 (85.8-99.9)	77 (32.5-95.9)	97.9 (95.9-99)	.885
High-grade dysplasia (n=19)	94.7 (73.9-99.8)	90.2 (78.5-96.7)	33.7 (18-54.9)	99.6 (97.9-99.9)	.857

BEDDLM, Barrett esophagus dysplasia deep learning model; NDBE, nondysplastic Barrett esophagus; WSI, whole slide imaging.

dysplasia grade on WSIs.[7-11] The Barrett esophagus dysplasia deep learning model is an ensemble deep learning model composed of an object detection model that makes a first-pass prediction and identifies regions of interest with bounding boxes, followed by a classifier model; the final label for each slide is assigned on the basis of the highest grade of agreement between the object detection and classifier models (Figure 2). Sensitivity, specificity, positive predictive value, and F1 score were reported and compared with the previously published results of the model. Of note, the used WSIs are the same slides that were used as the test set of the original model.

RESULTS

We analyzed 70 slides with a mean image size of 59,139 (SD, 15,998) × 147,483 (SD, 28,016) pixels. After running the developed algorithm on the slides, the mean image size was 40,831 (SD, 30,154) × 54,171 (SD, 46,283) pixels, resulting in a mean 7.11× (SD, 3.00) compression. Table summarizes the model performance with and without the size reduction algorithm.

DISCUSSION

We developed an image-processing algorithm that can reduce WSI size by removing non—tissue-containing background with an average of 7.11× size compression, without losing clinically relevant information as judged by similar model effectiveness when comparing the performance before and after reduction.

WSI is a digital method for capturing high-resolution images of tissue sections for pathology diagnosis. However, these images can be quite large, making it difficult to store, transmit, and process them efficiently. To address this issue, size reduction algorithms are often applied to WSI to reduce their file size. There are several algorithms that can be used for WSI size reduction, such as lossless compression and lossy compression.[18-23]

There are several algorithms that can be used for WSI size reduction, such as lossless compression and lossy compression. Lossless compression methods, such as Huffman coding or Run-Length Encoding, retain all information in the original image but reduce its size by efficiently encoding the data.[24,25] On the contrary, lossy compression methods, such as joint photographic experts group or wavelet compression, reduce the file size by discarding some of the image information. These methods can result in lower image quality but can achieve significant size reductions.[19] Nonetheless, joint photographic experts group compression can still be used as an additional compression technique in our pipeline as long as it remains below the threshold where diagnostic performance starts to decline.

Another approach to WSI size reduction is to downsample the image, which involves reducing the number of pixels in the image while preserving its overall shape and structure. This can be done by averaging or filtering the image or by selecting a subset of the original pixels. Downsampled images have smaller file sizes but may lose some of the diagnostic details in the process.[26]

Although there are existing tools that use similar methodologies to mask tissue-containing portions, our entire pipeline introduces a novel approach to compressing WSIs (Pocock et al. 2022).[27] Our proposed method saves the exact pixel value on tissue-containing parts of the WSI; in addition, the sized-reduced images do not affect the performance of downstream machine learning—based models. This has significant implications for clinicians and researchers, allowing for improved histopathology workflow in clinical and research pathology units, decreased need for storage space of WSIs, efficient downloading and transfer of images, standardized global educational possibilities, and increased collaborative opportunities for pathologists and researchers.

CONCLUSION

In conclusion, we have found that an automated tool that eliminates WSI background without removing the tissue-containing portion does not result in data loss, confirmed by similar performance when slides are interpreted by a deep learning model, which may aid clinicians and researchers in storing, transferring, and efficiently analyzing WSIs for use in their workflow.

POTENTIAL COMPETING INTERESTS

Dr Erickson is an officer at FlowSIGMA Inc and Yunu Inc. Dr Iyer reports research funding from Exact Sciences, Pentax Medical, and Cernostics; is a consultant for Exact Sciences, Medtronic, Ambu Medical, Symple Surgical, and Cernostics. Given their role as Editorial Board Member, Dr. Bradley Erickson, had no involvement in the peer-review of this article and has no access to information regarding its peer-review. All other authors disclosed no financial relationships.

Abbreviations and Acronyms: **BEDDLM,** Barrett esophagus dysplasia deep learning model; **HGD,** high-grade dysplasia; **LGD,** low-grade dysplasia; **NDBE,** nondysplastic Barrett esophagus; **RGB,** red-green-blue; **ROI,** regions of interest; **WSI,** whole slide imaging; **YOLO,** you only look once

Correspondence: Address to Bradley Erickson, MD, PhD, Mayo Clinic, 200 1st Street, SW, Rochester, MN 55905 (bje@mayo.edu).

ORCID

Shahriar Faghani: https://orcid.org/0000-0003-3275-2971|Bradley J. Erickson: https://orcid.org/0000-0001-7926-6095

REFERENCES

1. Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: whole-slide imaging and beyond. *Annu Rev Pathol*. 2013;8:331-359.
2. Montezuma D, Monteiro A, Fraga J, et al. Digital pathology implementation in private practice: specific challenges and opportunities. *Diagnostics (Basel)*. 2022;12(2):529. https://doi.org/10.3390/diagnostics12020529.
3. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology*. 2012;61(1):1-9.
4. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. Posted online March 3, 2017. arXiv [csCV]. http://arxiv.org/abs/1703.02442.
5. Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab Invest*. 2021;101(4):412-422.
6. Krempel R, Kulkarni P, Yim A, et al. Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB). *BMC Bioinformatics*. 2018;19(1):156.
7. Faghani S, Codipilly DC, Vogelsang D, et al. Development of a deep learning model for the histologic diagnosis of dysplasia in Barrett's esophagus. *Gastrointest Endosc*. 2022;96(6):918-925.e3.
8. Hsu WW, Guo JM, Pei L, et al. A weakly supervised deep learning-based method for glioma subtype classification using WSI and mpMRIs. *Sci Rep*. 2022;12(1):6111.
9. Wu Y, Cheng M, Huang S, et al. Recent advances of deep learning for computational histopathology: principles and applications. *Cancers (Basel)*. 2022;14(5):1199. https://doi.org/10.3390/cancers14051199.
10. Singhal N, Soni S, Bonthu S, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Sci Rep*. 2022;12(1):3383.
11. Abdel-Nabi H, Ali M, Awajan A, et al. A comprehensive review of the deep learning-based tumor analysis approaches in histopathological images: segmentation, classification and multi-learning tasks. *Clust Comput*. Published online January 9, 2023. https://doi.org/10.1007/s10586-022-03951-2.
12. Yagi Y, Gilbertson JR. Digital imaging in pathology: the case for standardization. *J Telemed Telecare*. 2005;11(3):109-116.
13. Weinstein RS, Graham AR, Richter LC, et al. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. *Hum Pathol*. 2009;40(8):1057-1069.
14. Gohlke C. *Cgohlke/tifffile: v2022.5.4* 2022. https://doi.org/10.5281/zenodo.6795861.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *arXiv [csLG]*. 2012;85:2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Accessed January 31, 2023.
16. Cacchiani V, Iori M, Locatelli A, Martello S. Knapsack problems—an overview of recent advances. Part II: multiple, multidimensional, and quadratic knapsack problems. *Comput Oper Res*. 2022;143:105693.
17. Andersson D. Rectangle-packer: rectangle packing program. Github. https://github.com/Penlect/rectangle-packer. Accessed January 31, 2023.
18. Isola J. Optimal image data compression for whole slide images. *Diagn Pathol*. 2016;1(8). https://doi.org/10.17629/www.diagnosticpathology.eu-2016-8:172.
19. Krupinski EA, Johnson JP, Jaw S, Graham AR, Weinstein RS. Compressing pathology whole-slide images using a human and model observer evaluation. *J Pathol Inform*. 2012;3:17.
20. Zarella MD, Jakubowski J. Video compression to support the expansion of whole-slide imaging into cytology. *J Med Imaging (Bellingham)*. 2019;6(4):047502.
21. DICOM whole slide imaging. https://dicom.nema.org/dicom/dicomwsi/. Accessed February 1, 2023.

22. Hulsken B. Fast compression method for medical images on the web. Posted online May 18, 2020. arXiv [eessIV], http://arxiv.org/abs/2005.08713.
23. Sharma A, Bautista P, Yagi Y. Balancing image quality and compression factor for special stains whole slide images. *Anal Cell Pathol (Amst)*. 2012;35(2):101-106.
24. Lakhani G. Modified JPEG Huffman coding. *IEEE Trans Image Process*. 2003;12(2):159-169.
25. Bradley SD. Optimizing a scheme for run length encoding. *Proc IEEE*. 1969;57(1):108-109.
26. Zarella MD, Bowman D, Aeffner F, et al. A practical guide to whole slide imaging: a white paper from the Digital Pathology Association. *Arch Pathol Lab Med*. 2019;143(2):222-234.
27. Pocock J, Graham S, Vu QD, et al. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Commun Med*. 2022;2:120.