Digital Pathology with the Gastric Histopathology Sub-size Image Database (GasHisSDB)

Applying machine learning methods to IHC images has been performed in several cancers (breast(1), prostate(2), stomach(3)). However, its often the case that the datasets used in the model training are not publicly available(3) and thus not fully reproducible or transparent for other researchers. A 2022 paper by Hu et al. describes the curation of the first publicly available dataset for gastric histopathology images, specifically for the goal of allowing researchers to improve on machine learning methods to detect these types of cancer. Here, we replicated the results found in the paper by independently training 2 deep learning models using the GasHisSDB dataset and test the model using whole slide images (WSI) of adenocarcinoma from the TCGA-STAD repository.

## Current Digital Pathology Challenges

The current gold standard for many cancer type diagnosis is based on visual inspection of tissue sections by a trained pathologist(4). While the accuracy of trained pathologists with adequate time is close to 100%, these conditions are not often met due to resource issues. The 2017 Royal college of Pathologist survey found that only 3% of departments have enough stay and 45% have had to send work away to outside labs for processing.

Diagram

Description automatically generated with medium confidence

Figure 2 - 2017 Royal College of Pathologists Survey of 103 Histopathology Departments. Sourced here(5).

## A.I.s potential for Digital Pathology

Current performance of A.I. models to distinguish non-neoplastic tissue from adenoma and adenocarcinomas is comparable to pathologists in research settings can reach accuracies over 95%(3). The potential of A.I. model goes beyond classification, with papers demonstrating models capable of

1. Identifying treatment response to immunotherapies in melanoma(6) and lung cancers(7)
2. Predicting TP53 mutation status across numerous cancer types(8)
3. Predicting survival rates across numerous cancer types(9)

WSI provide a great resource for A.I. models as they typically store far more information than other types of medical imagery (CT scans(10) for example). However, all of these examples are still in the research phase and have not been incorporates into clinical settings yet. The only FDA approached A.I. pathology software is used for prostate tumor detection, Paige Prostate Detect(11).

Chart, radar chart

Description automatically generatedChart, radar chart

Description automatically generated

Figure 3 - Applications of DL models in pathology. Taken from here(10). References refer to those in the original paper.

## GasHisSDB Paper Replication

### Dataset creation - Patching

The authors(12) took 600 H&E stained gastric whole slide images (WSI) (2048 x 2048 pixels) and cropped them into 3 collection patches of different sizes (160 x 160, 120 x 120, 80 x 80) according to a defined process. The patching process is necessary as the pre-trained CNN models are not designed to work with the original larger images.

Diagram

Description automatically generatedFigure 4 - Patch Creation Process for dataset

The patching & labelling process is:

1. WSI were assessed by 6 pathologists from Longhua Hospital, Shang-hai, China.
2. Any areas highlighted as abnormal were noted as the ground truth for the classification algorithms to be used (a)
3. WSI that contain any abnormal areas were cropped into the 3 sizes (b)
4. Patches were then filtered so that only those patches with > 50% abnormal area were considered with the subset dataset. (c)
5. WSI considered fully normal were cropped into the 3 sizes. (d)
6. Images were pre-processed by random rotation to reduce correlation amongst images. (e)

The processed dataset was downloaded from [gitee](https://gitee.com/neuhwm/GasHisSDB.git.) repository provided by the authors (gitee appears to be a Chinese version of GitHub). The dataset was not already split into train, validation, and test sub-sets so I performed this operation myself in a 4:3:3 ratio. The 160 subset dataset contains 13,124 abnormal and 20,160 normal images, giving a class balance of approximately 40/60 Abnormal-to-Normal.

### Applying Transfer Learning to DP

The GasHisSDB paper uses a machine-learning technique called transfer learning. This is a technique where models that have been trained on one task have been retrained to perform another task. Adjusting an existing model allows the new model to benefit from the patterns already learned in the previous model. The paper uses pre-trained image classification models that have been trained to classify images from the ImageNet dataset. The ImageNet dataset is a large collection of images which have been labelled with 1 of 1000 different classifications, for example:

* “Pembroke Welsh corgi”
* “cocktail shaker”
* “hen-of-the-woods” (a type of mushroom)

The ImageNet dataset was curated to serve as the basis for the ImageNet Challenge, a competition where participants attempt to classify an image with 1 of 1000 labels. Milestones of computer vision innovations were often denoted by jumps in performance in this competition. The problem was declared solved in 2015 with the ResNet model as it achieved super-human results.

Chart, bar chart

Description automatically generated

Figure 5 - ImageNet Challenge Results. 2012 is the 1st deep learning model, AlexNet.

The deep learning model outputs used in the challenge output 1000 probability values (1 for each class) and the top-5 highest probability is chosen as the predicted labels. To classify the WSI as Abnormal/Normal, we need to adjust the last layer of the model to produce just 1 probability, the probability of being Abnormal say, instead of 1000.

### Building a Deep Learning Model using the GasHisSDB dataset

In the GasHisSDB paper, they use 3 architectures: VGGNet, ResNet and a more recent transformer model. The transformer model did not achieve the convergence in the same training time as the VGGNet/ResNet models, so I focused on replicating just the first 2 model results. The paper also trained the models are the 3 patch sizes that they generated. I focused on just the 160x160 images as the principles employed at the same for the other 2 sizes. The authors also did not provide any code for the models they created. This is not an insurmountable problem as the process for creating such a model is straight forward in principle. However, as we will see, it puts unnecessary ambiguity into how the reported results were achieved.

Calendar

Description automatically generated

Figure 6 - Curated datasets

### Training Results on GasHisSDB dataset

Training the 2 architectures on a training and validation test set produced some interesting results. The number of epochs used (100) was the same as the original authors. Google Colab GPU resources were used to speed up the training process. The VGG model began to overfit the training data after just 12 epochs. Another interest point to note is that both models achieved quite reasonable results after just 2 epochs, with VGG having 88% accuracy and ResNet 75%. These values should be compared against the base line class balance of 60% (meaning a model that just returns “Normal” for all images would achieve a 60% accuracy score).

Chart, histogram

Description automatically generatedChart

Description automatically generated with medium confidence

Figure 7 – Training Results. VGG overfit after just 12 epochs but performed better than ResNet which is surprising.

Comparing the best epoch training results of both models against each other: VGG achieved 90.46% accuracy after 12 and ResNet achieved 88.6% after 100.

### Test Results on GasHisSDB dataset

True model performance can only be measured on unseen data however. Typically, if a model has been trained appropriately so that it generalisesgeneralizes well, you would expect test accuracy to be similar but lower than the validation accuracy results. Using the test dataset, the results were counter intuitive as VGG the test accuracy was higher. The VGG model, which we saw was also overfitting the data implying poor generalization, achieved 94.12% accuracy. ResNet achieved 88.5417%, a more intuitive figure (lower but similar). This points to possible issues with the splitting of the data into the subsets, such that the distribution of images in the test set was not the same as the distribution for those in the train or validation set.

### Performance Results on TCGA dataset

While the results of the classifier were high with the unseen GasHisSDB data, a true real world performance would be on datasets outside of those curated by the original authors. The TCGA STAD project has WSI for over 379 patients with adenocarcinomas(13). The full dataset of images is over 320GB is size. A sample of 40 of the smaller diagnostic scans were downloaded and patches of 160x160 were created using modified code modified from a GitHub repository(14). This code script utilizes the open source OpenSlide software to read WSI for processing. The patching code attempts to screen each patch to determine if there is tissue present or not, based on a saturation threshold image processing technique.

One immediate issue with using the TCGA dataset is that we do not have tissue level/pixel level annotations, only a single whole-slide level label. This means for slides that have both normal and abnormal sections, patches will have the wrong label. To test this dataset using our model architecture, we must assume that the any 160x160 image from this dataset would be determined to be cancerous by a trained pathologist. This is not a viable assumption, but it is required without more resources (e.g. a pathologist, more annotations etc. better training model process).

Using the patching script, patches were created using 40 WSI samples. Of those patches 35 were randomly sampled for evaluation with our trained classifier. Using the VGGNet model (as it performed the best), the model only correctly classified 15/35 of the samples as abnormal or 42.86%. This is worse than a random flip of a coin.

A picture containing clothing, bedclothes, fabric

Description automatically generated

Figure 8 – Patches created using our modified script (25/36 patches that were tested)

## Overview of the paper and our replication

As we’ve seen, the problems faced by A.I. applications in WSI goes well beyond just classification accuracy.

### Generalization

Our model was able to achieve 90%+ accuracy on the original curated dataset, but poorly on the “general” TCGA set. There were several issues with how the general set was tested and these most certainly contributed to the poor results materially. There is also the issue of the homogeneity of the original 600 WSI images used to train the model. Since these were all collected from one institute they are likely to be less varied than would be expected in practice.

### WSI size

The original paper used WSI images that were just 2048 x 2048 pixels. It’s not uncommon for WSI images to be 100,000 x 100,000. This is 2x orders of magnitude larger. It’s likely the 2048x2048 images had a magnification or some other processing applied before the researchers applied the patching process.

### Annotations

Of the 2 publicly available datasets explored, there were issues with annotation. In the GasHisSDB, the ground truth WSI annotations (i.e. the original label areas of the original WSI that the trained pathologists identified) were not made available so that they could be investigated further.

In the TCGA dataset, there are no tissue level annotations only global labels, leading to a non-viable option of assume the entire slide is abnormal in nature.

### “Black-box” Model Effect

More broadly, the VGGNet and ResNet models (Convolutional Neural Network, CNN, designs) suffer from a “black-box” model effect, whereby its not clear what the models are identifying in the images to determine the output classification. The models employed only return labels and probabilities. They do not explain clearly what features of the images they are using to discern the labels. A famous example how these black-box models can misbehave was a model that was used to classify skin lesions. Further analysis of how the model worked revealed that it heavily relied on the presence of a ruler to determine if it was a malignancy or not. The model had learned that images with rulers are more likely to malignancy and was using this information increase its accuracy scores(15).

## Summary

In short, the GasHisSDB highlights the need for further publicly available datasets of whole slide images for cancer detection. Our replication results were broadly in line with those seen in the original paper, and we demonstrated the potential failure of such models when brought outside the controlled confines of the research setting and into the “real world”.

A [Google Colab notebook](https://colab.research.google.com/drive/14px49cJBeCkJQZ_HlNPlCIvGYNSCuKVZ?usp=sharing) is available for reference, and all code and training results can be replicated from that notebook.

# References

1. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Scientific Reports. 2017;7(1):46450.

2. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21(2):233-41.

3. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. Scientific Reports. 2020;10(1):1504.

4. Elmore J. Abstract SY01-03: The gold standard cancer diagnosis: Studies of physician variability, interpretive behavior, and the impact of AI. Cancer Research. 2021;81(13\_Supplement):SY01-3-SY-3.

5. Pathologists RCo. Royal College of Pathologists Survey Results 2017. In: RCP, editor. 2017.

6. Harder N, Schönmeyer R, Nekolla K, Meier A, Brieu N, Vanegas C, et al. Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. Scientific Reports. 2019;9(1):7449.

7. Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Cǎlin AM, Delli Pizzi A, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol. 2019;30(6):998-1004.

8. Noorbakhsh J, Farahmand S, Foroughi pour A, Namburi S, Caruana D, Rimm D, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nature Communications. 2020;11(1):6367.

9. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLOS ONE. 2020;15(6):e0233678.

10. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. British Journal of Cancer. 2021;124(4):686-96.

11. Paige Prostate Detect 2022 [Available from: <https://paige.ai/clinical/>.

12. Hu W, Li C, Li X, Rahaman MM, Ma J, Zhang Y, et al. GasHisSDB: A New Gastric Histopathology Image Dataset for Computer Aid ed Diagnosis of Gastric Cancer: arXiv; 2021 2021.

13. Lucchesi FR, & Aredes, N. D. . Radiology Data from The Cancer Genome Atlas Stomach Adenocarcinoma [TCGA-STAD] collection. The Cancer Imaging Archive. . 2016.

14. TSRAI. WSI\_patches GitHub Repo. 2020.

15. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. Journal of Investigative Dermatology. 2018;138(10):2108-10.

Dylan’s References: I will edit \*\*

Note: 1, 2, 3 in lindsay’s risk factor table – MTHFR polymorphisms

1.

Han Z, Sheng H, Gao Q, Fan Y, Xie X. Associations of the MTHFR rs1801133 polymorphism with gastric cancer risk in the Chinese Han population. Biomedical Reports. 2020 Nov 17;14(1).

2.

Petrone I, Bernardo PS, dos Santos EC, Abdelhay E. MTHFR C677T and A1298C Polymorphisms in Breast Cancer, Gliomas and Gastric Cancer: A Review. Genes. 2021 Apr 17;12(4):587.

3.

CHEN L, LU N, ZHANG B-H, WENG L, LU J. Association between the MTHFR C677T polymorphism and gastric cancer susceptibility: A meta-analysis of 5,757 cases and 8,501 controls. Oncology Letters. 2015 Jun 10;10(2):1159–65.

4.

Jahn SW, Plass M, Moinfar F. Digital Pathology: Advantages, Limitations and Emerging Perspectives. Journal of Clinical Medicine [Internet]. 2020 Nov 18;9(11):3697. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7698715/

5.

Betmouni S. Diagnostic digital pathology implementation: Learning from the digital health experience. DIGITAL HEALTH. 2021 Jan;7:205520762110202.

6.

Montezuma D, Monteiro A, Fraga J, Ribeiro L, Gonçalves S, Tavares A, et al. Digital Pathology Implementation in Private Practice: Specific Challenges and Opportunities. Diagnostics. 2022 Feb 18;12(2):529.

7.

Cui M, Zhang DY. Artificial intelligence and computational pathology. Laboratory Investigation. 2021 Jan 16;101(4):412–22.

8.

Pell R, Oien K, Robinson M, Pitman H, Rajpoot N, Rittscher J, et al. The use of digital pathology and image analysis in clinical trials. The Journal of Pathology: Clinical Research. 2019 Mar 25;5(2):81–90.

9.

Hanna MG, Parwani A, Sirintrapun SJ. Whole Slide Imaging. Advances in Anatomic Pathology. 2020 Jul;27(4):251–9.

10.

Vodovnik A. Diagnostic time in digital pathology: A comparative study on 400 cases. Journal of Pathology Informatics. 2016;7(1):4.

11.

Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Modern Pathology. 2021 Oct 5;35.

12.

Clarke E, Doherty D, Randell R, Grek J, Thomas R, Ruddle RA, et al. Faster than light (microscopy): superiority of digital pathology over microscopy for assessment of immunohistochemistry. Journal of Clinical Pathology [Internet]. 2022 Jan 15 [cited 2022 Mar 5]; Available from: https://jcp.bmj.com/content/early/2022/01/16/jclinpath-2021-207961

13.

Lahiani A, Gildenblat J, Klaman I, Albarqouni S, Navab N, Klaiman E. Virtualization of Tissue Staining in Digital Pathology Using an Unsupervised Deep Learning Approach. Digital Pathology. 2019;47–55.

14.

de Haan K, Zhang Y, Zuckerman JE, Liu T, Sisk AE, Diaz MFP, et al. Deep learning-based transformation of H&E stained tissues into special stains. Nature Communications. 2021 Aug 12;12(1).

15.

Magaki S, Hojat SA, Wei B, So A, Yong WH. An Introduction to the Performance of Immunohistochemistry. Methods in Molecular Biology (Clifton, NJ) [Internet]. 2019;1897:289–98. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30539453

16.

Bell D. Haematoxylin and eosin stain. Radiopaediaorg. 2020 May 11;

17.

Identifying Cells And Epithelia Lab [Internet]. medcell.org. Medcell; [cited 2022 Apr 19]. Available from: http://medcell.org/tbl/identifying\_cells\_and\_epithelia/reading.php

18.

Li Y, Li N, Yu X, Huang K, Zheng T, Cheng X, et al. Hematoxylin and eosin staining of intact tissues via delipidation and ultrasound. Scientific Reports [Internet]. 2018 Aug 16;8(1):1–8. Available from: https://www.nature.com/articles/s41598-018-30755-5

19.

Sui W, Ou M, Chen J, Wan Y, Peng H, Qi M, et al. Comparison of immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) assessment for Her-2 status in breast cancer. World Journal of Surgical Oncology. 2009;7(1):83.

20.

PubChem. Rhodamine B [Internet]. pubchem. PubChem; 2020. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Rhodamine-B

21.

Young RJ, Möller A. Immunohistochemical Detection of Tumour Hypoxia. Histology Protocols [Internet]. 2009 Nov 9 [cited 2022 Apr 19];151–9. Available from: https://www.researchgate.net/publication/40447323\_Immunohistochemical\_Detection\_of\_Tumour\_Hypoxia

22.

File:HSP IF IgA.jpg - Wikipedia [Internet]. commons.wikimedia.org. 2008 [cited 2022 Apr 19]. Available from: https://en.wikipedia.org/wiki/File:HSP\_IF\_IgA.jpg

23.

Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nature Communications. 2020 Aug 27;11(1).

24.

Wang S, Zhu Y, Yu L, Chen H, Lin H, Wan X, et al. RMDL: Recalibrated Multi-instance Deep Learning for Whole Slide Gastric Image Classification. Medical Image Analysis. 2019 Aug;101549.

25.

Wang X, Chen Y, Gao Y, Zhang H, Guan Z, Dong Z, et al. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nature Communications. 2021 Mar 12;12(1).

Kevin’s references

1. Shen Bo ZJ. Research progress on blood supply of gastric cancer and its arterial interventional chemotherapy. World Journal of Chinese Digestion. 2003.

2. Wieduwilt MJ, Moasser MM. The epidermal growth factor receptor family: biology driving targeted therapeutics. Cell Mol Life Sci. 2008;65(10):1566-84.

3. Gang Niu XC. Vascular Endothelial Growth Factor as an Anti-angiogenic Target for Cancer Therapy. Curr Drug Targets. 2010.

4. <https://www.cancer.org/cancer/stomach-cancer/treating/targeted-therapies.html>. Targeted Drug Therapy for Stomach Cancer.

5. <https://www.cancer.org/content/dam/CRC/PDF/Public/8841.00.pdf>. Treating Stomach Cancer.

6. Cheng G, Mei Y, Pan X, Liu M, Wu S. Expression of HER2/c-erbB-2, EGFR Protein in Gastric Carcinoma and its Clinical Significance. Open Life Sci. 2019;14:119-25.

7. Fisusi FA, Akala EO. Drug Combinations in Breast Cancer Therapy. Pharm Nanotechnol. 2019;7(1):3-23.