OHSU Goecks Lab: Leveraging Publicly Available Imaging Data and Machine Learning to Predict Clinical Subtypes of Breast Cancer

Sophia S., Davin M., Sam K., Pranav M.

Background:

A common clinical method for breast cancer subtype diagnosis are pathological annotations of H&E images. Correct subtype identification enables medical professionals to customize breast cancer treatments based on patient profile and better predict survival outcomes. These annotations currently require a trained pathologist to manually scan slides of differentially-dyed nuclei and cytoplasm to identify specific cellular formations characteristic of known breast cancer histological (ILC/IDC) and PAM50 subtypes. While this manual diagnostic method is extremely accurate, it is also slow and labor-intensive which prevents wide-spread implementation in clinical settings. Therefore, the goal of our project is to apply image pre-processing, deep learning, and machine learning techniques to demonstrate the feasibility of accurately automating breast cancer subtype categorization of clinical H&E images in the TCGA database. If successful, this automation could eventually allow for quicker diagnosis of breast cancer subtypes and lead to more effective, targeted therapeutics.

Software:

1. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature*. **585**, 357–362 (2020).

Original paper for NumPy, a mathematical library for Python. We will use this to remove image tiles at the ends of the images and with low tissue content using means and standard deviations. As well as other data analysis and visualization tools. https://numpy.org/doc/stable/

2. A. Goode, B. Gilbert, J. Harkes, D. Jukic, M. Satyanarayanan, OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*. **4**, 27 (2013).

Original paper for OpenSlide, a python library used to work with SVS images in Python. We will use this to open, view, and tile SVS image objects in Python. https://openslide.org/api/python/

3. K. Inc, histomicstk: A Python toolkit for Histopathology Image Analysis, (available at https://github.com/DigitalSlideArchive/HistomicsTK).

A Python package used for pathology image analysis that we may use to normalize stains between H&E slide images. https://pypi.org/project/histomicstk/

4. StainTools — StainTools documentation, (available at https://staintools.readthedocs.io/en/latest/).

A python package used for stain normalization that we may use to normalize or augment stain between different H&E slide images. https://staintools.readthedocs.io/en/latest/

5. A. Paszke, S. Gross, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library, 12 (2019).

Original paper publishing PyTorch, a Python library for deep learning that we are using to design, train, and validate our ML model to identify breast cancer subtypes from H&E stain images. PyTorch 1.12 documentation available at https://pytorch.org/docs/stable/index.html.

6. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*. **9**, 90–95 (2007).

Original paper for Matplotlib, a Python package used for data visualization. We will mostly use this for data exploration and code testing, though some of the final images summarizing our results may use this package as well. https://matplotlib.org/stable/index.html

7. Overview — CuPy 11.2.0 documentation, (available at https://docs.cupy.dev/en/stable/overview.html).

A python package to incorporate Compute Unified Device Architecture (CUDA) for processes in the NumPy and SciPy. This package will replace NumPy in almost all code, except where it is not compatible with other Python packages. The authors of this package have shown that for many common mathematical functions, CuPy may be up to an order of magnitude faster than NumPy.

8. Scikit-learn documentation — scikit-learn 1.1.2 documentation, (available at

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning).

Scikit-learn is a python package that includes many scientific and mathematical tools. For this project we are using it to fit gaussian distributions to a histogram of mean pixel intensity to find a cutoff to filter tiles without stain. This method is called Gaussian Mixture Modeling. Time permitting, it may also be useful to compare a second machine learning algorithm to our Pytorch model.

Mentor Articles:

9. The Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature*. **490**, 61–70 (2012).

Original TCGA-BRCA publication describing results of implementing several different machine learning models - trained to identify PAM50 subtypes - to find common genomic,proteomic, and/or clinical features characteristic to these breast cancer subtypes (Luminal A, Luminal B, Basal-like, HER2E). This paper provides background on our project subject and examples of previously applied methods.

10. G. Ciriello, M. L. Gatza, A. H. Beck, et al., Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*. **163**, 506–519 (2015).

Secondary TCGA-BRCA publication where double the sample size of the original 2012 article was used with this machine learning method to distinguish any key genetic traits defining ILC and IDC breast cancer subtypes, and also focuses on how studying these differentially expressed genes could lead to improved therapeutic cancer treatments. This paper provides additional examples of applicable ML models, applied to larger sample sizes.

11. J. S. Parker, M. Mullins, M. C. U. Cheang, et al., Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *JCO*. **27**, 1160–1167 (2009).

Original article publishing the developmental results of the PAM50 subtype-differentiating machine learning method that assigns standardized classifications: Luminal A, Luminal B, Basal-like, HER2E (enriched), and normal-like (control). Also includes graphical predictions of untreated and treated survival probabilities based on breast cancer subtype. This paper describes each of the PAM50 breast cancer subtypes our ML model will be able to differentiate between.

12. J. van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic. *Nat Med*. **27**, 775–784 (2021).

A review paper summarizing the history of computational pathology (CPATH), the initial development and recent improvements to convolutional neural networks (CNNs) in clinical settings, current challenges to wide-scale CPATH implementation, and potential future directions of this field. This review provided us with a general idea of past accomplishments and current challenges in the fields of deep learning and ML.

13. J. Saltz, R. Gupta, L. Hou, et al., Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Reports*. **23**, 181-193.e7 (2018).

Example implementation of TGCA H&E image analysis conducted by CNN deep learning methods (computational staining) to collect data on breast cancer cell structural patterns specific to differing tumor-infiltrating lymphocyte (TIL) statuses. The patterns/features identified with the described computational method are also compared to manual pathologist annotations to assess method accuracy. This paper contains more examples of ML algorithms applied to H&E images (albeit with different subtype classifications), as well as discussion on how ML methods complement/compare to manual pathologist annotations.

14. S. Farahmand, A. I. Fernandez, F. S. Ahmed, et al., Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. *Mod Pathol*. **35**, 44–51 (2022).

Example implementation of using “transfer-learning” and “full-learning” CNNs to identify HER2 subtypes (+/-/other). Different annotating, image pre-processing, and machine learning methods are compared to assess the most pathologically accurate combination of methods for HER2 breast cancer subtype identification. Also describes how developed models can be used to predict survival/response to treatment (trastuzumab: HER2 expression blocker). This paper compares accuracy results between transfer learning and custom CNN models, which can help us decide which to use in our project.

15. H. D. Couture, L. A. Williams, J. Geradts, et al., Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast Cancer*. **4**, 30 (2018).

Article focused on assessing pathological accuracy of a transfer learning model (VGG16 CNN) to identify and categorize CBCS3 H&E images by breast cancer severity grade (low-intermediate, high), ER status (+, -, unknown), histological subtype (ILC, IDC), and PAM50 subtype (Luminal A, Luminal B, Basal-like, HER2, Normal-like, unknown). From these characteristics, risk of recurrence (ROR) is also calculated. Strengths and

weaknesses of the applied transfer learning method in this application are discussed, which will help us determine if using a pre-trained model is possible for our project.

16. M. Macenko, M. Niethammer, J. S. Marron, et al., "A method for normalizing histology slides for quantitative analysis" in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (IEEE, Boston, MA, USA, 2009; http://ieeexplore.ieee.org/document/5193250/), pp. 1107–1110.

A white paper describing a suggested method for H&E image pre-processing, along with key method algorithms and example test set (melanoma) results. Steps include stain variation correction and intensity variation correction. Results show that the documented pre-processing image adjustments allow for more accurate and conclusive differentiation in downstream deep learning analysis. This paper has guided development of our own pre-processing steps for this project.

Other Articles:

17. D. Bychkov, N. Linder, A. Tiulpin, H. Kücükel, M. Lundin, S. Nordling, H. Sihto, J. Isola, T. Lehtimäki, P.-L. Kellokumpu-Lehtinen, K. von Smitten, H. Joensuu, J. Lundin, Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Sci Rep*. **11**, 4037 (2021).

This is a study from 2021 that uses gene expression of the gene *ERBB2* to weakly supervise (uses an additional dataset to inform the training of a machine model) a deep learning model analyzing H&E stains to identify cancer *ERBB2* status. We found this article by footnote chasing from “Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer” by Farahmand et al.. They found that this method was able to identify features to predict *ERBB2* gene amplification and from this inform effective treatment. This paper could be useful as an extremely recent study using many of the same methods that we might. In particular, it outlines the use of transfer learning in a CNN using PyTorch, something we still need to explore.

18. A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*. **7**, 29 (2016).

This paper serves as a “tutorial” for using deep learning as a tool for image analysis. It covers general steps in the process as well as specific use cases and some programs. This paper is from 2016 and thus some of the specific programs and other details are out of date, the general workflow could be useful to us as we work on our project. It is highly

cited (953 citations on Google Scholar) and may be particularly useful for working out the math of image input size into our neural network.

19. Lin, Yuqi et al. “Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data.” *Genes* vol. 11,8 888. 4 Aug. 2020, doi:10.3390/genes11080888

This paper classifies breast cancer subtypes using multi-omics data from the cancer genome atlas. They use deep neural networks and can give us ideas to adapt into our imaging classification model. This was a decently cited paper in Web of Science, with 19 citations and 47 references. This paper is particularly useful for getting a better idea of how people have used deep learning neural networks to classify other breast cancer subtypes using a different format of data.

20. M. Veta, J. P. W. Pluim, P. J. van Diest and M. A. Viergever, "Breast Cancer Histopathology Image Analysis: A Review," in *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400-1411, May 2014, doi: 10.1109/TBME.2014.2303852.

This paper provides a general overview of histopathology and breast cancer. Image details, and H&E stain examples are provided in section III - staining analysis. This was highly cited in Web of Science (338 citations) and is a top result when searching for “Breast Cancer Histopathology” when filtering by highly cited papers. This paper is particularly useful for background on hematoxylin and eosin stains, and what components of their images are useful for categorization.

21. Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, A. María Vanegas, Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models. *Sensors*. **20**, 4373 (2020).

This paper has detailed preprocessing and training sections for a similar dataset to what our project will deal with. This study also utilizes learning models for image classification, and their citations may be a useful reference as we follow our workflow. This paper was found through footnote chasing a highly cited review paper overviewing CNN architectures and applications (346 citations on web of science) in order to find a more specific reference for our application. This is useful to our project as it provides an example workflow and approach which aligns with our goals.

22. K. Venetis, E. Crimini, E. Sajjadi, C. Corti, E. Guerini-Rocco, G. Viale, G. Curigliano, C. Criscitiello, N. Fusco, HER2 Low, Ultra-low, and Novel Complementary Biomarkers: Expanding the Spectrum of HER2 Positivity in Breast Cancer. *Front. Mol. Biosci.* **9**, 834651 (2022).

We found this article by looking at the most relevant cited article that used the Farahmand et al., 2022 article (pearl growing on Web of Science). This one - 7 citations and 95 cited references - was the most applicable to our project goal of using ML algorithms to identify breast cancer subtypes based on H&E images, as it discussed the difficulties that current pathological methods - both manual and computational - are encountering in clinical settings. Recognizing these challenges will help us customize our model to best address current “gaps” in computational H&E image annotation methods.

23. M. Salvi, U. R. Acharya, F. Molinari, K. M. Meiburger, The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*. **128**, 104129 (2021).

We found this article by looking at the most relevant cited article that used the Couture et al., 2018 article (pearl growing on Web of Science). This one - 42 citations and 203 cited references - was the most applicable to our project goal of using ML algorithms to identify breast cancer subtypes based on H&E images, as it described how optimizing pre- and post-processing steps applied to each image can help increase model accuracy. We can reference this paper as we are evaluating our own ML model to determine the necessity of adding additional pre-processing steps as a means of increasing model subtype identification ability.

24. S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, N. Navab, AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Transactions on Medical Imaging*. **35**, 1313–1321 (2016).

This article studies the use of crowd-sourced pre-trained CNN models to address the lack of available “ground-truth” data in medical imaging. We found it through footnote chasing from Deep learning in histopathology: the path to the clinic (van der laak et al.), but it is also highly cited (306 citations on Web of Science). Though our method is considerably different from theirs, Albarqouni et al. describe the math behind the machine learning algorithm they use, which could inform how we create our neural network(s).

25. L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data. 8, 53 (2021).

This review paper discusses broad applications of convolutional neural networks, and mentions their use in classifying H&E stained images. It is useful to our project as a starting point for footnote chasing an example of this application of CNNs to data which aligns with ours. We used this source to footnote chase (21, Hameed et al.) , and was found whilst searching highly cited papers on Web of Science, which deal with the application of CNNs.