The Impact of Digital Histopathology Batch Effect on Deep Learning Model Accuracy and Bias

Frederick M. Howard[1], James Dolezal[1], Sara Kochanny[1], Jefree Schulte[2], Heather Chen[2], Lara Heij[3,4], Dezheng Huo[5], Rita Nanda[1], Olufunmilayo I. Olopade[1], Jakob N. Kather[6,7,8], Robert Grossman[1], Alexander T. Pearson[1]

[1] Department of Medicine, University of Chicago, Chicago, IL, USA

[2] Department of Pathology, University of Chicago, Chicago, IL, USA

[3] Department of Surgery and Transplantation, University Hospital RWTH Aachen, Aachen, Germany

[4] Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany

[5] Department of Public Health Sciences, University of Chicago, Chicago, IL, USA.

[6] Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany

[7] Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom

[8] Medical Oncology, National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg, Germany

Abstract

The Cancer Genome Atlas (TCGA) is one of the largest biorepositories of digital histology. Deep learning (DL) models have been trained on TCGA to predict numerous features directly from histology, including survival, gene expression patterns, and driver mutations. However, we demonstrate that these features vary substantially across tissue submitting sites in TCGA for over 3,000 patients with six cancer subtypes. Additionally, we show that image differences between submitting sites can easily be identified with DL. This site detection remains possible despite commonly used color normalization and augmentation methods, and we quantify the digital image characteristics constituting this histologic batch effect. As an example, we show that patient ethnicity within the TCGA breast cancer cohort can be inferred from histology due to site-level batch effect, which must be accounted for to ensure equitable application of DL. Batch effect also leads to overoptimistic estimates of model performance, and we propose a quadratic programming method to guide validation that abrogates this bias.

Main

A standard component of the diagnosis of nearly all human cancers is the histologic examination of hematoxylin and eosin stained tumor biopsy sections. Histologic features identified by pathologists help characterize tumor subtypes, prognosis[1], and at times can predict response to treatment. Quantification of more subtle pathologic features can further discriminate between good and poor prognosis tumors, such as the quantification of tumor infiltrating lymphocytes in breast cancer, but such detailed analysis can be time consuming and variable between pathologists[2]. The increasing availability of digital histology coupled with advances in artificial intelligence and image recognition has led to computational approaches to rigorously assess pathologic features associated with a variety of tumor specific factors. Deep learning is a subdomain of artificial intelligence, referring to the use of multilayer neural networks to identify increasingly higher-order features of images to allow for accurate classification.

Deep learning on digital histology has exploded as a potential tool to identify standard histologic features such as grade[3,4], mitosis,[5,6] and invasion[7,8]. Recently, deep learning approaches have been applied to identify less apparent histologic features, including clinical biomarkers such as breast cancer receptor status[4,9], microsatellite instability[10,11], or the presence of pathogenic virus in cancer[12]. These approaches have been further extended to infer more subtle features of disease, including gene expression[13–15] and pathogenic mutations[16,17]. The predictive accuracy of many of these models have been validated in external datasets, but many studies rely on single data sources for both training and validation.

The Cancer Genome Atlas (TCGA) has been critical for development of deep learning histology models, containing over 20,000 digital slide images from 24 tumor types, along with associated clinical, genomic, and radiomic data[18]. Due to the propensity of machine learning algorithms to overfit, performance is typically reported in a reserved testing set or evaluated with cross-validation, to avoid biased estimates of accuracy[19]. However, the propensity for overfitting of digital histology models to site level characteristics has been incompletely characterized and is infrequently accounted for in internal validation of deep learning models. The genomic batch effects in TCGA and other high throughput sequencing endeavors have been well characterized, and are the product of the hundreds of tissue source sites contributing samples and the multiple sites for genome sequencing and characterization[20–22]. However, histologic imaging data is also prone to batch effect, which may be unique from each tissue submitting site. Prior to sectioning, tissue is first fresh frozen or fixed in formalin and embedded in

paraffin, and each fixation method generates unique artifacts[23]. Slides are then stained with the eponymous hematoxylin and eosin stains, the color and intensity of which can vary based on the specific stain formulation and the amount of time each stain is applied. The digitization of slides may then vary due to scanner calibration and choice of resolution and magnification[24].

Several methods have been proposed to mitigate staining differences between slides[25], including methods designed for color variation across images by Reinhard and colleaegues[26], and methods designed specifically for histology by Macenko and colleagues[27]. Color augmentation (Figure 1), where the color channels of images are altered at random during training to prevent a model from learning stain characteristics of a specific site have also been utilized in histology deep learning tasks[28,29]. Most assessments of stain normalization and augmentation techniques have focused on the performance of models in validation sets, rather than true elimination of batch effect[30,31]. Here, we describe the clinical and slide level variability between sites in TCGA, and methods to ensure robust use of internal and external validation to minimize false positive findings with deep learning image analysis.

Results

Characterization of Clinical and Histologic Batch Effect in TCGA

Important clinical variables differ across tissue submitting sites across TCGA. It has been recognized previously that outcomes and survival vary across site for a number of cancers[32], but even more fundamental factors differ depending on submitting organization. We compared the distribution of basic demographics such as age, race, gender, and body weight index and tumor specific factors such as stage and histologic subtype. Sites were included for comparison if they submitted at least 20 tissue slides. For breast cancer (BRCA TCGA cohort), all demographic characteristics as well as estrogen receptor status (n = 969), progesterone receptor status (n = 969), HER2 expression (n= 847), PAM50 subtype (n = 914), BRCA1 mutational status (n = 931), immune subtype (n = 1,002), and 3-year progression free survival (n = 458)[33] varied significantly between cohorts, with false discovery rate correction and $P < 0.05$ (Figure 2). We systematically applied this approach to five other major solid tumor types, and demonstrate that multiple impactful clinical features vary by site for all tumor subtypes tested - including body mass index in colorectal cancer (colon adenocarcinoma (COAD) and rectal adenocarcinoma (READ) TCGA cohorts, n = 476), ALK translocation status in squamous cell lung cancer (LUSC TCGA cohort, n = 157) and lung adenocarcinoma (LUAD TCGA cohort, n = 118), and human papilloma virus (HPV) status in head and neck squamous cell carcinoma (HNSC TCGA cohort, n = 598) – all with $P < 0.05$ and significant after FDR correction; Supplemental Table 1). Of note given the increasing interest in developing survival models based on pathology, stage varied by site in all cancer subsets tested, and 3 year progression free survival varied across site in all cancers except lung adenocarcinoma.

We then applied classical descriptive statistics for image analysis to document the differences in slide features across site, calculating first order statistics and second order Haralick texture features for comparison across sites[34,35]. All first and second order statistics demonstrated variance according to tissue submitting site (Figure 3, Supplemental Table 2). Applying stain normalization techniques at a slide level for breast cancer improved some first order characteristics but measures of dissimilarity for all second order characteristics (as measured by F-statistic) remained greater than that of any first order characteristics (Figure 4). Of note, the second order feature angular second moment remained the most dissimilar feature (highest F-statistic) with any form of stain normalization for all cancer types except lung and head and neck squamous cell carcinoma (Supplemental Table 2).

Deep Learning Algorithms Accurately Identify Tissue Submitting Site

To assess the ability of deep learning to predict tissue submitting site, we trained a deep learning convolutional neural network to predict site. To assess accuracy of site prediction, we used 3-fold cross validation stratified by site (Figure 5a), and calculated the one versus rest area under the receive operating characteristic (AUROC) curve (Supplemental Table 2). The features used by such a model to predict site can be illustrated with a UMAP[36] representation of final layer activations, with representative slide tiles selected for each UMAP coordinate – in this case demonstrated a clear basophilic-eosinophilic color gradient on the X-axis (Figure 5b). To assess the ability of stain normalization and color augmentation to prevent prediction of site, we repeated this process with normalization or augmentation applied at the tile level for the six examined cancer subtypes (Supplemental Table 3). Site discrimination was highly accurate at baseline, with an average one-versus-rest (OVR) area under the receiver operating characteristic curve (AUROC) ranging from 0.998 for clear cell renal cancer (TCGA-KIRC, n = 513) to 0.966 for TCGA-HNSC (n = 450). Stain normalization techniques modestly decreased accuracy of site prediction, but remained highly accurate with an average OVR AUROC of over 0.850 with all normalization techniques for all cancers. In five of the six cancers, one form of grayscale normalization was able to significantly reduce the site prediction compared to baseline as assessed by paired two-sided t-test.

Naturally, if a deep learning model can distinguish sites based on non-biologic differences between slide staining patterns and slide acquisition techniques, models will learn the clinical variability between samples at different sites. This is analogous to the Husky versus Wolf problem, where a deep learning model falsely distinguishes pictures of these two canines based on the fact that more wolves are pictured in snow[37]. To demonstrate this fact, we train deep learning models to predict ethnicity, which is highly variable across sites, using two different methods of cross validation (Figure 5c). First, stratifying by race and ignoring site, race can be accurately predicted (Supplemental Table 4). We can correct for biased results by ensuring sites are isolated to single data fold, or **preserved site** cross validation. However, it is still necessary to stratify by the variable of interest, as stratification improves the variance and bias of estimates of accuracy[38]. It is not possible to perfectly stratify by race, but we can select partitions of sites as close as possible for k-fold cross-validation using quadratic programming[39].

This methodology produces perfect stratification (defined as differences by no more than 1 for the number of patients in each category between generated data folds) for preserved site cross validation for all clinical variables described in Figure 2 in the breast cancer dataset, except for gene expression subtype. Notably, estimated AUROC for the prediction of ethnicity in TCGA-BRCA (n = 913) is significantly ($p < 0.001$ for all comparisons) greater than random chance (which would equate to an AUROC of 0.5) for all stain normalization methods when stratifying by race, but with preserved site cross validation, no model suggests an accuracy greater than random chance (Supplemental Table 4).

We can further demonstrate that models are weighting slide staining characteristics in decision making by examining the predictions for tissue submitting site OL / University of Chicago, the only site where patients of African ancestry comprises the majority of samples. For patients in the validation data folds, false positive predictions for African ancestry (measured at the tile level, n = 2,206) are significantly higher for standard cross validation balanced by race, as compared to preserved site cross validation (Figure 6, Supplemental Table 5). In other words, standard cross validation inaccurately classifies European patients from a site with predominant African ancestry, as the decision is likely related to site

level image characteristics. Conversely, the difference in rates of false positive European ancestry is less pronounced, and high rates of false positivity in both models are likely due to the predominance of European patients across training data.

Discussion

We have demonstrated that batch effect exists in the histology images in TCGA across multiple cancer types, and inadequately controlling for this batch effect results in biased estimates of accuracy. Although stain normalization can remove some of the perceptible variation and augmentation can mask differences in color, second order image features are unaffected by these methods, and they do not resolve the ability of deep learning models to accurately identify a tissue submitting site.

Multiple studies have trained and validated models using cases from TCGA without external validation or isolating sites to either the training or validation datasets. Such studies include the classification of cancer histology in lung cancer[17], genetic mutation prediction in multiple cancer types[16,17,40], prediction of grade in clear cell renal cancer[41], prediction of breast cancer molecular subtype[42], the prediction of gene expression[13], or correlation of histology and outcome[40,43]. Survival outcomes are particularly challenging to develop rigorous models for using histology from TCGA, and model performance may be falsely elevated not only by the disparate outcomes across sites, but also the site level differences in critical factors relevant to survival such as stage and age. Studies demonstrating histologic discrimination of survival and recurrence in glioblastoma[41,43,44], renal cell cancer[41], and lung cancer[45] patients from TCGA which lack external validation cohorts may have biased estimates of outcome. Prediction of survival may also suffer from this bias[46] even when correcting for age, stage, and sex, as other factors that vary by site also contribute to outcome, ranging from ethnicity of enrollees, to the treatment available at academic vs community centers. Given that traditional image and textural features vary between sites in TCGA, it is likely that non-deep learning prognostic studies that predict prognosis from traditional image analysis features may suffer from a similar bias[47]. Certainly, there are numerous models which perform well in TCGA without site isolation, and continue to accurately predict outcomes of interest with external validation cohorts, such as prediction of microsatellite instability or BRAF mutations in colon cancer[11,16]. A number of prognostic models have demonstrated the ability of deep learning, trained on a diverse cohort, to accurately predict prognosis in cancers such as colorectal cancer[48,49] and mesothelioma[50]. However, until the external validity of findings is verified, studies that do not specify the split of sites in their validation and test cohorts must be looked at with a certain degree of scrutiny.

We have also demonstrated that deep learning models have the potential to predict ethnicity from histology *directly* from learned patterns about the demographic makeup of tissue submitting sites. This poses a challenging ethical dilemma for the implementation of deep learning histology models. It has been well documented that black women with breast cancer have a poorer prognosis that is not completely accounted for by stage and receptor subtype[51]. Contributing factors may include delays in treatment initiation and inadequate intensity of therapy[52]. As histology models may be able to identify ethnicity by proxy, they must be carefully implemented in an equitable fashion to avoid inappropriate conclusions[53]. For example, if a prognostic model has learned the staining pattern of a site that recruits primarily African patients with a high rate of recurrence unrelated to disease biology, it may denote low risk patients at that site as 'poor prognosis', which could lead to inappropriate administration of chemotherapy or other treatments.

Best Practices for Addressing Batch Effect for Deep Learning on Histology

We recommend a series of best practices for deep learning studies on histology using TCGA or other combined datasets of multiple hospital sites. First, the variation of outcomes of interest should be reported across included sites. This will allow an assessment of the potential impact histologic batch effect can have on accuracy. Additionally, knowledge about the distribution of outcomes on the training and testing sites can allow for accurate assessment of model performance, as AUROC is an uninformative marker for heavily imbalanced datasets, where the precision recall curve can be more informative[54]. This will also allow an estimate of the impact of batch effect on model predictions. Even if performance stands the test of external validation, models may retain the biases learned from institutional staining patterns. Thus, if outcomes of interest vary heavily across sites, further prospective validation at individual institutions may be necessary before implementation.

If variation of outcomes are seen across sites, models should not be trained and assessed for accuracy on patients from the same contributing site. As we have demonstrated, including sites within the validation and training datasets results in biased estimates of accuracy. The tried and true gold standard for any artificial intelligence endeavor is external validation, which also ensures that not only site level but dataset level batch effects are not reflected in model performance[55]. However, adequate external validation datasets are not frequently unavailable, and it is important to accurately assess the promise of models at an early stage before significant time is spent in further research and investigation. It is rarely possible to perfectly stratify patients by the outcome of interest when isolating sites to either validation or training data, but stratification is still important for unbiased estimates of accuracy. We propose using convex optimization / quadratic programming to identify the split of sites across to allow optimal stratification. This can also be applied to linear outcomes by stratifying the outcome of interest into meaningful subgroups or quartiles prior to optimization.

Finally, stain normalization and color augmentation techniques should still be used to improve model accuracy in external validation and implementation. Although normalization and augmentation do not completely prevent models from learning site specific characteristics, several studies have reported greater validation accuracies with the use of such techniques[30,31]. It is likely that these techniques eliminate some but not all of the reliance that deep learning models have on stain related features; by making the differences in slide characteristics more subtle, models may be more likely to pick up on biologically relevant factors. Grayscale normalization likely discards some relevant biologic information[56], but the ideal method for normalization has not yet been definitively determined, and may vary per application of interest. Similarly, although attempting to normalize second order features derived from the gray level co-occurrence matrix may render sites more indistinguishable, such features are closely associated with intrinsic tumor biology and must likely be preserved for deep learning applications[57].

Limitations

We present a comprehensive description of pixel level characteristics across site in TCGA using classical image analysis techniques, however other factors likely contribute to the detectable differences between sites. Although the compression factor and resolution also varies between the slide files obtained at different sites, we did not assess the ability of deep learning models to detect these factors. It is likely that other higher order image features contribute to the site level differences, such as Gabor, wavelet packet, and multiwavelet features.[24] However, extensive characterization of all described

textural features is not necessary to demonstrate the presence of batch effect and the impact this has on model performance.

We have chosen a subset of proposed stain correction methods, but there have been innumerable approaches and variations that may further reduce the intra-site variability in staining. An unsupervised learning approach to normalizing stains has been proposed, but did not outperform methods of augmentation without normalization in test datasets[30]. Adversarial networks may also allow for models to avoid learning undesirable characteristics of datasets[58].

Conclusions

Digital histology in TCGA carries a multifactorial signature that is characteristic of the tissue submitting site. This signature can be easily identified by deep learning models, and can lead to an overestimation of model accuracy, if hospital sites are included in both the training and validation datasets. Care should be taken to describe the batch effect of outcomes of interest across sites, and if significant, a submitting site should be isolated to either the cohort used for training or for testing a model. A quadratic programming approach can maintain optimal stratification while still isolating submitting sites to either training or validation datasets.

Methods

Patient cohorts

Patient data and whole slide images were selected from 6 of the tumor types from TCGA with the highest number of slides available to better identify site to site batch effect. Tumor types included breast (BRCA)[59], colorectal (COAD and READ – with data combined for sites enrolling to both cohorts)[60], lung squamous cell carcinoma (LUSC)[61], lung adenocarcinoma (LUAD)[62], renal clear cell (KIRC)[63], and head and neck squamous cell carcinoma (HNSC)[64]. Slides and associated clinical data was accessed through the Genomic Data Commons Portal (https://portal.gdc.cancer.gov/). Ancestry was determined using genomic ancestry calls as per Carrot-Zhang and colleagues[65]. Immune subtypes were used from the work published by Thorsson and colleagues[33]. Informed consent was obtained for all participants in TCGA.

Image Processing and Deep Learning Model

Scanned whole slide images of hematoxylin and eosin stained tissue were acquired in SVS format. Each slide was reviewed by a pathologist for manual annotation of area of tumor, to ensure ink or other non-cancer artifacts did not influence slide level statistics[31]. For slide level first order and second order statistical analysis, slides were downsampled to 5 microns per pixel, or approximately 2X magnification. For deep learning applications, the tumor region of interest is tessalated into 299 pixel x 299 pixel tiles for evaluation, each representing a 302 µm x 302 µm area of histology, effectively generating an optical magnification of 10X. Convolutional neural network models are written in Python 3.8 using TensorFlow and structured based on Xception[66] with pretraining weights used from ImageNet[67]. This is followed by a single hidden layer with width 500 and a softmax layer for prediction. Further details regarding implementation have been previously published[16,19].

Each tile is assigned a label associated with the outcome of interest. Tile libraries were also balanced by category to eliminate bias, such that the number of tiles for each target category was equivalent.

Models are trained to categorical outcomes using the Adam optimizer and sparse categorical crossentropy loss. Stain normalization and augmentation is applied to individual tiles at the time of training and assessment. Macenko and Reinhard normalization is applied as previously described[26,27], grayscale refers to direct slide conversion to grayscale, and 'grayscale normalized' refers to conversion to grayscale with histogram equalization[68]. Both light and heavy levels of hue saturation value (HSV) augmentation was applied, with light augmentation multiplying each of these three channels by a scalar from 0.9 to 1.1, and heavy augmentation multiplying the hue and saturation channels by a random scalar from 0.7 to 1.3. Additionally, further augmentation through random tile rotation is performed, and further normalization ensures inputs have a mean of zero and variance of one. Models are trained with 3-fold cross validation, learning from two splits of the data and then evaluated on the third split (Figure 5). Models were trained on 3 epochs of the training dataset. Deep learning model training and evaluation was performed on 16 deep learning-specific NVidia Tesla V100s graphical processing unit (GPU) nodes within a HIPAA-compliant environment.

Statistics

To quantify differences between categorical clinical features across sites, a Chi-squared test is used, with significance determined using a false discovery rate (FDR) of 0.05 with the Benjamini-Hochberg method. A similar approach is applied using ANOVA to identify differences between slide level characteristics between sites, with the same FDR. First order statistics are calculated from individual red, green, and blue pixel values across images, and include mean, standard deviation, skewness, kurtosis, and entropy, the latter being calculated as follows:

$$Skewness = \frac{m_3}{{m_2}^{3/2}} \quad (1)$$

$$Kurtosis = \frac{m_4}{{m_2}^{2}} \quad (2)$$

Where $m_i = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^i$

$$Entropy = -\sum_{n=1}^{N} x_n * \log(x_n) \quad (3)$$

Second order Haralick features[35] were calculated from the gray level co-occurrence matrix *P*:

$$Contrast = \sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2 \quad (4)$$

$$Dissimilarity = \sum_{i,j=0}^{levels-1} P_{i,j}|i-j| \quad (5)$$

$$Homogeneity = \sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2} \quad (6)$$

$$Angular\ Smooth\ Momentum = \sum_{i,j=0}^{levels-1} P_{i,j}^2 \quad (7)$$

$$Correlation = \sum_{i,j=0}^{levels-1} P_{i,j} \frac{(i-u_i)(j-u_j)}{\sqrt{o_i^2\ x\ o_j^2}} \quad (8)$$

Similar values for calculated features were seen for angles of 0°, 45°, 90°, and 135° so reported values for second order features are averaged across these four angles. The AUROC values are reported as

means and ranges generated with 3-fold cross validation. Comparisons between average AUROC for prediction of site with different methods of stain normalization was performed with a two sided paired t-test with an FDR of 0.05. To calculate k-folds for cross-validation where each site is isolated to a single fold, we define the following convex optimization problem. If $m_{s,c}$ is a binary variable indicating if site s is a member of fold c, and $n_{s,f}$ is a integer indicating the number of samples from site in the categorical feature class f, then we seek to minimize the mean squared error of divergence from perfect stratification:

$$Error = \sum_{\substack{f \in \textbf{Features}, \\ \textbf{c} \in \textbf{Crossfolds}}} \left(\sum_{s \in \textbf{Sites}} m_{s,c} \cdot n_{s,f} - \frac{\sum_{s \in \textbf{Sites}} n_{s,f}}{k} \right)^2 \tag{9}$$

With the constraints that for all sites s:

$$\sum_{c \in \textbf{Crossfolds}} m_{s,c} = 1 \tag{10}$$

We used CPLEX v12.10, IBM to solve the optimal solution of equations (9) and (10)[69]. Our code used for fold generation for cross-validation is available from https://github.com/fmhoward/PreservedSiteCV. Comparisons between AUROC values and random chance (0.50) was performed using a Z-test with an FDR of 0.05. Comparisons between the false positive rates for African ancestry were performed using a chi-squared test at a tile level.

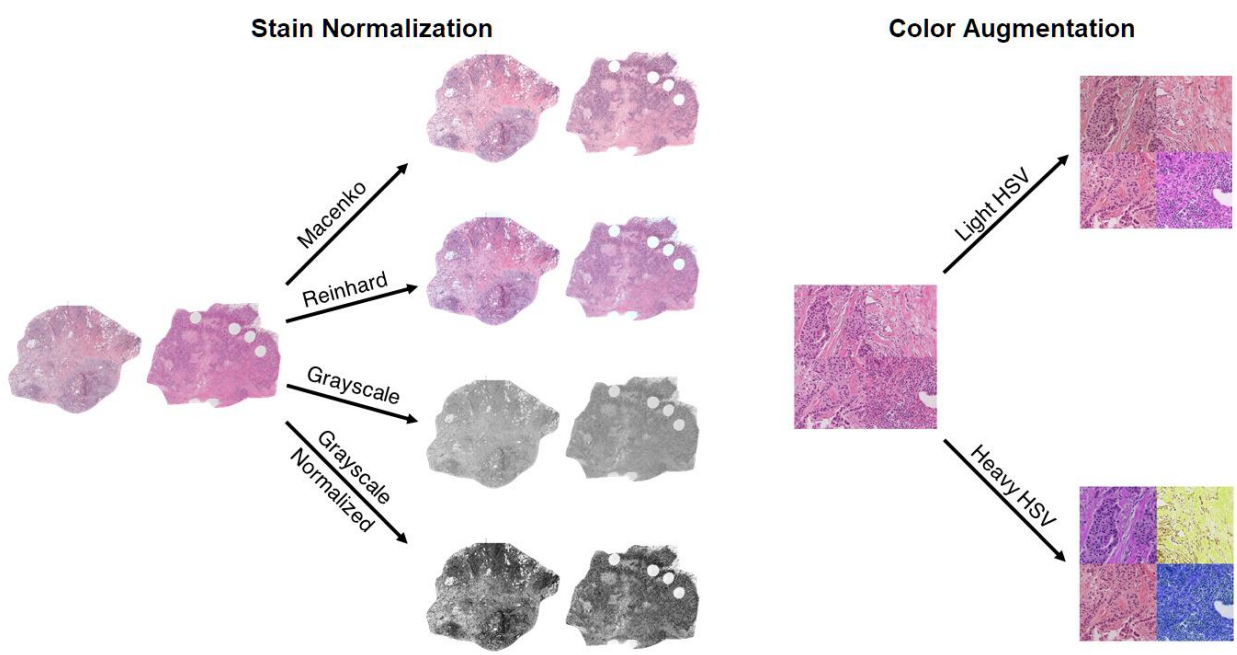

Figure 1. Previously proposed methods of stain normalization and augmentation. Stain normalization refers to changes in color characteristics to reduce the effect of staining differences between slides.

Augmentation refers to random variations applied to individual tiles during machine learning to prevent overfitting with regards to the varied characteristic.

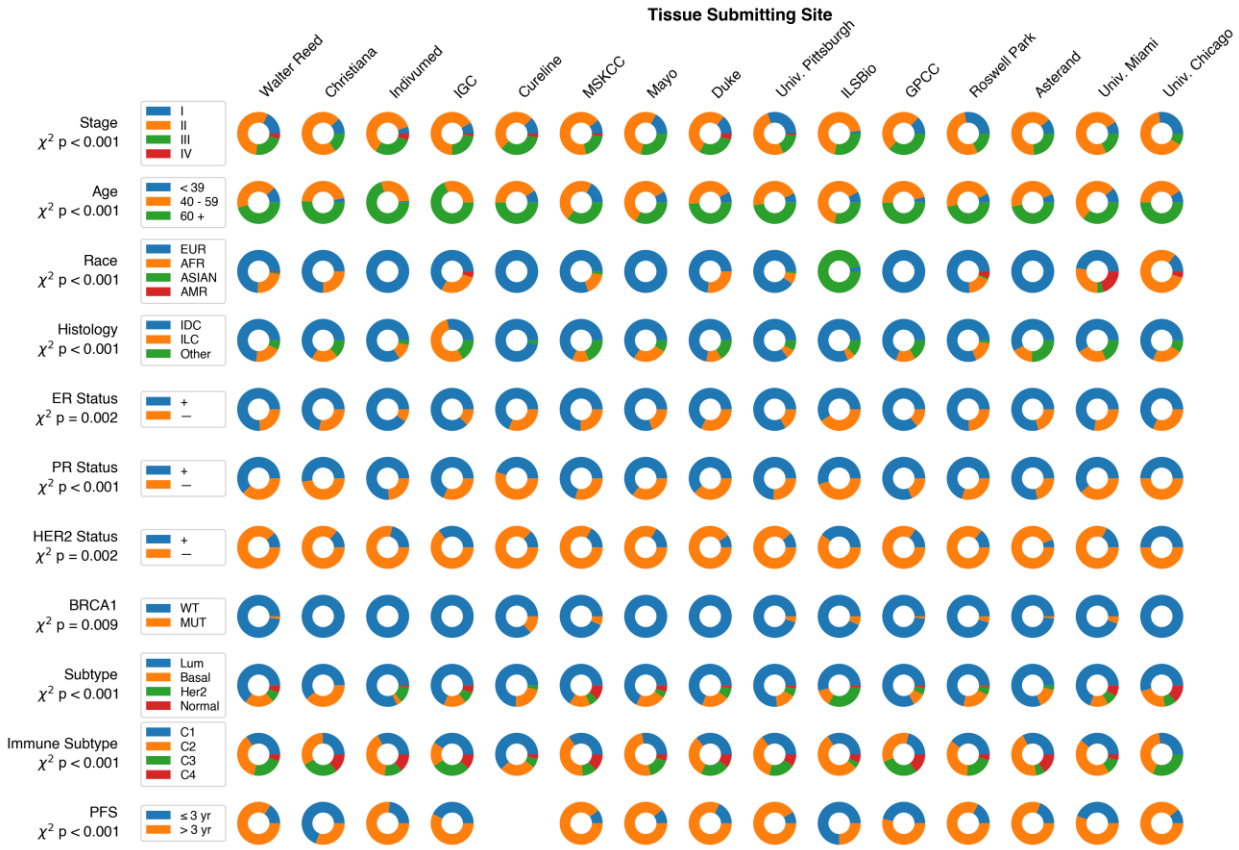

Figure 2. Demographics and Tumor Characteristics of Breast Cancer across Sites with 20 or more samples in TCGA.

Abbreviations: IGC = International Genomics Consortium. MSKCC = Memorial Sloan Kettering Cancer Center. GPCC = Greater Poland Cancer Center. EUR = European, AFR = African, AMR = Native American, IDC = Invasive Ductal Carcinoma, ILC = Invasive Lobular Carcinoma. WT = Wild type. MUT = Mutant. Lum = Luminal. PFS = Progression Free Survival.

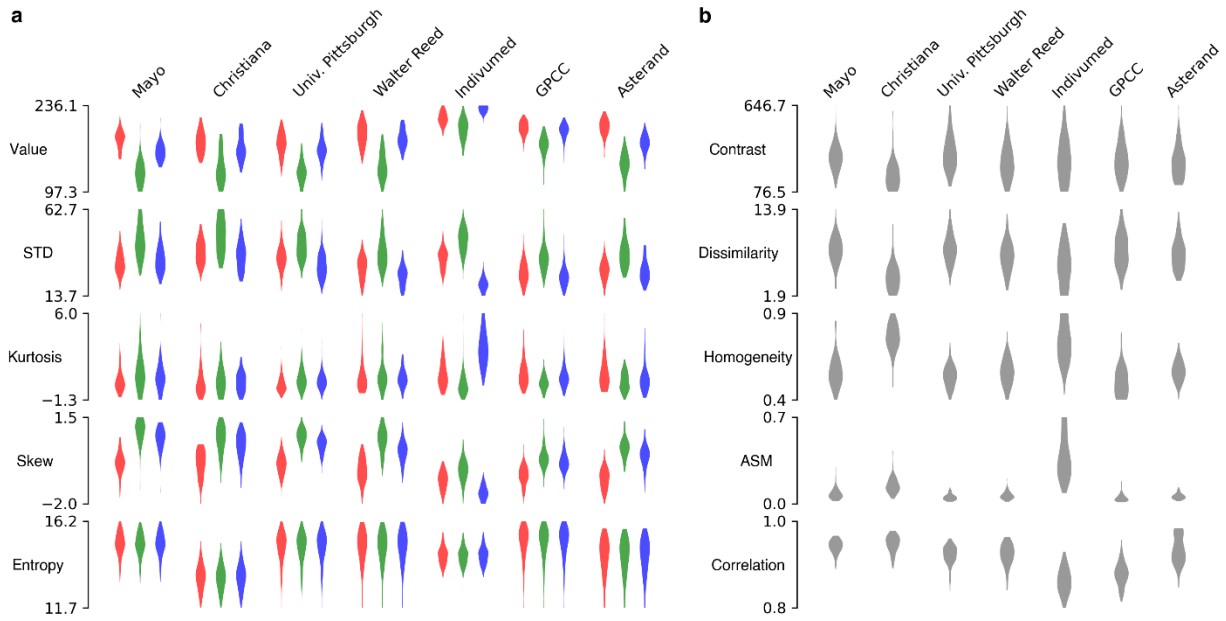

Figure 3. Breast Cancer Digital Histology Feature Variability in TCGA. Sites contributing at least 50 slides are included, demonstrating that image feature variation is not solely a function of small sites that infrequently contributed to TCGA. **a.** First order characteristics for red, green, and blue are shown in their respective colors. **b.** Haralick second order textural features also vary by submitting site. **Abbreviations: STD = Standard Deviation, ASM = Angular Second Moment. GPCC = Greater Poland Cancer Center.**

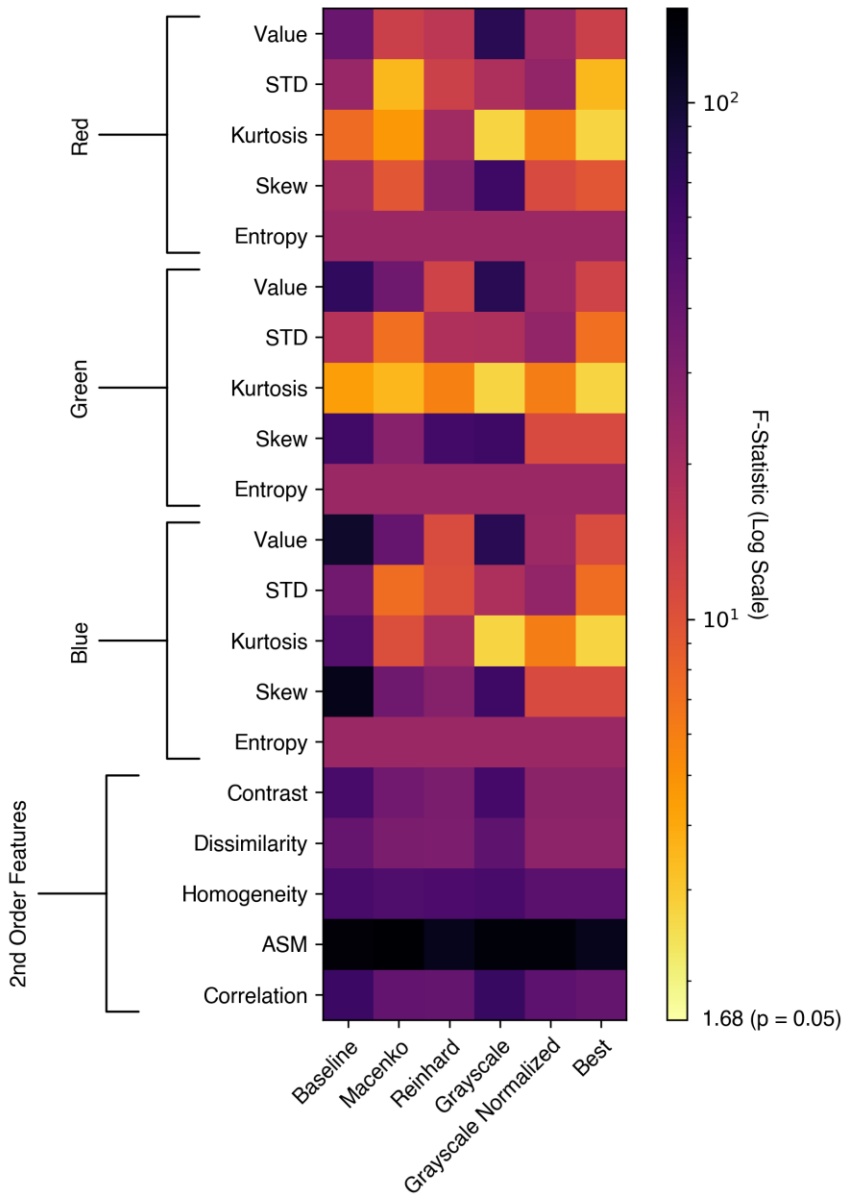

Figure 4. ANOVA F-Statistic for First and Second Order Image Features, TCGA-BRCA. Stain normalization does not resolve first order stain variability (by F-statistic), and minimal impact is seen on second order features.

Abbreviations: STD = Standard Deviation; ASM = Angular Second Moment

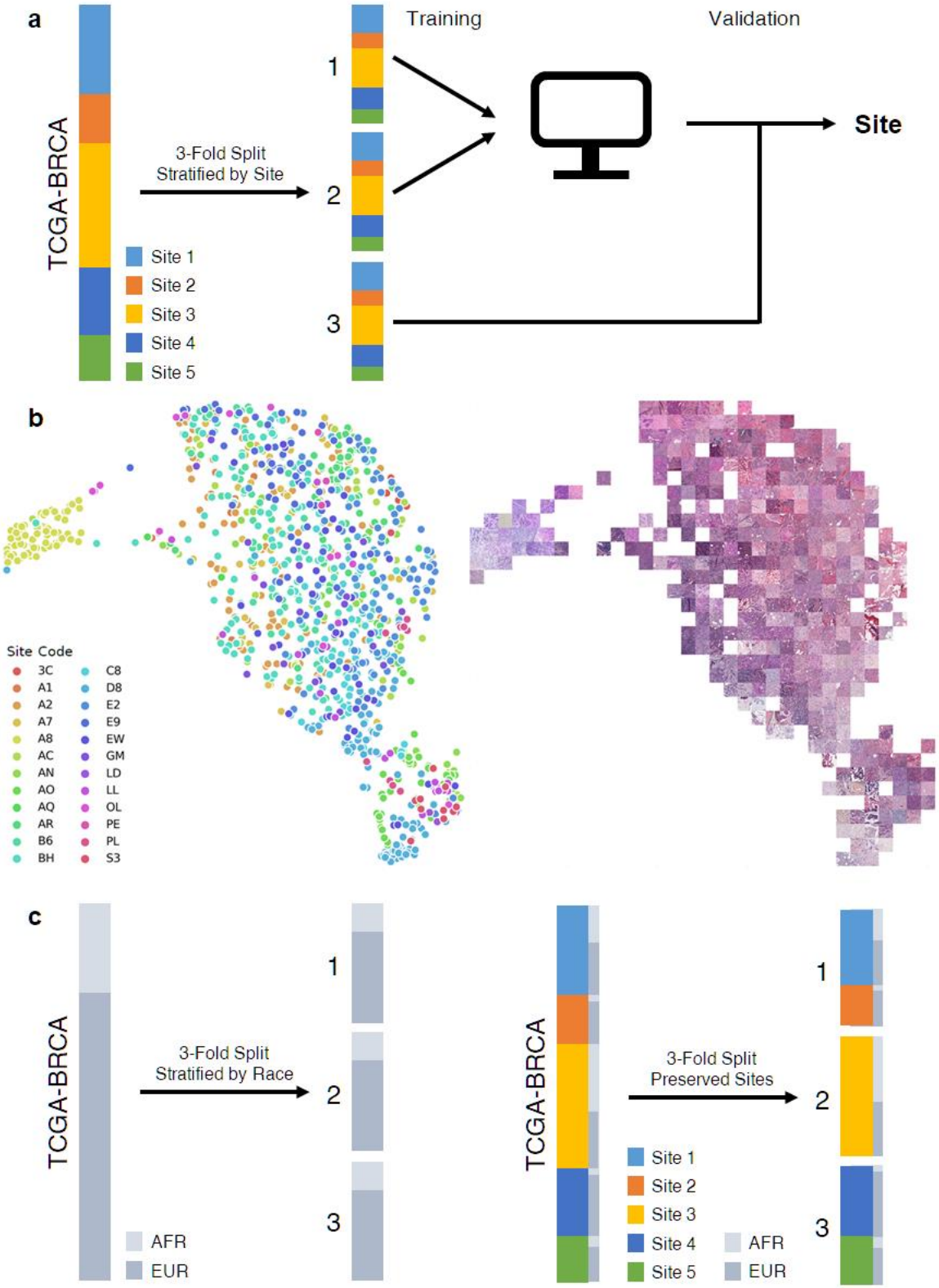

Figure 5. a. Setup for cross validation for prediction of tissue submitting site. Data is split into three folds, with each site represented equally in all folds. Cross validation is then performed, where a model

is trained on two of the datasets and performance is assessed on the third dataset. This process is repeated threefold for an averaged performance metric. **b.** UMAP representation of final activation of model trained to recognize submitting site. Each point on the left figure represents the centroid tile from a single slide. The nearest tile to each UMAP coordinate is visualized on the right. **c.** We assess the impact of including slides from a tissue submitting site within both the training and validation sets on the prediction of race, using two methods of generating folds for cross validation. First, we split the data into three folds, stratifying by race, irrespective of site. For a comparator, we split the data into three folds where each site is isolated into a single fold, with the secondary objective of equalizing the ratio African and European ancestry in each fold.

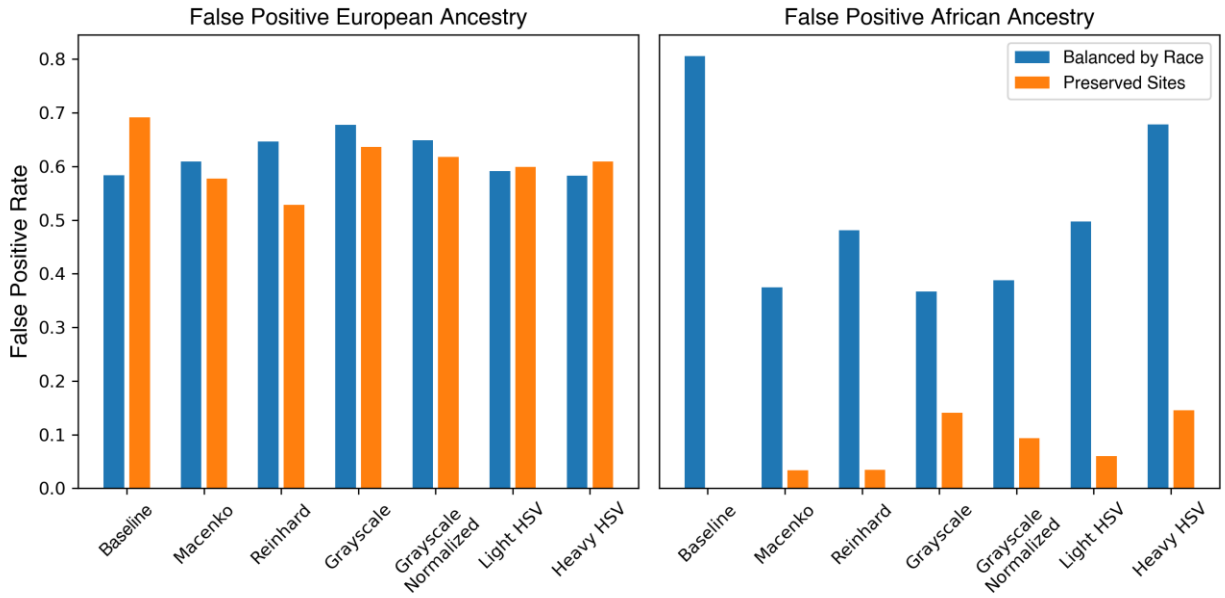

Figure 6. False Positive Rates for Prediction of Ancestry, University of Chicago site. The University of Chicago site submitted a predominance of African ancestry samples to TCGA, and we assess the performance of predictions when models are balanced by race (and University of Chicago patients are thus included in all three folds used for cross-validation), and when sites are isolated to a single fold. False positive rates are similar for European ancestry, but dramatically higher for African ancestry, indicating that models are learning the slide characteristics of samples from University of Chicago and using this information to make calls on patient ethnicity.

1. Carriaga MT, Henson DE. The histologic grading of cancer. *Cancer*. 1995;75(1 Suppl):406-421. doi:10.1002/1097-0142(19950101)75:1+<406::aid-cncr2820751322>3.0.co;2-w

2. Salgado R, Denkert C, Demaria S, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann Oncol*. 2015;26(2):259-271. doi:10.1093/annonc/mdu450

3. Bulten W, Pinckaers H, Boven H van, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*. 2020;21(2):233-241. doi:10.1016/S1470-2045(19)30739-9

4. Couture HD, Williams LA, Geradts J, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast Cancer*. 2018;4(1):1-8. doi:10.1038/s41523-018-0079-1

5. Saha M, Chakraborty C, Racoceanu D. Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Medical Imaging and Graphics*. 2018;64:29-40. doi:10.1016/j.compmedimag.2017.12.001

6. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Trans Med Imaging*. 2016;35(5):1313-1321. doi:10.1109/TMI.2016.2528120

7. 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. doi:10.1038/s41598-020-58467-9

8. Cruz-Roa A, Gilmore H, Basavanhally A, 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. doi:10.1038/srep46450

9. Shamai G, Binenbaum Y, Slossberg R, Duek I, Gil Z, Kimmel R. Artificial Intelligence Algorithms to Assess Hormonal Status From Tissue Microarrays in Patients With Breast Cancer. *JAMA Netw Open*. 2019;2(7). doi:10.1001/jamanetworkopen.2019.7700

10. Echle A, Grabsch HI, Quirke P, et al. Clinical-grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology*. Published online June 17, 2020. doi:10.1053/j.gastro.2020.06.021

11. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*. 2019;25(7):1054-1056. doi:10.1038/s41591-019-0462-y

12. Kather JN, Schulte J, Grabsch HI, et al. Deep learning detects virus presence in cancer histology. *bioRxiv*. Published online July 5, 2019:690206. doi:10.1101/690206

13. Schmauch B, Romagnoni A, Pronier E, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications*. 2020;11(1):3877. doi:10.1038/s41467-020-17678-4

14. He B, Bergenstråhle L, Stenbeck L, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*. 2020;4(8):827-834. doi:10.1038/s41551-020-0578-x

15. Badea L, Stănescu E. Identifying Transcriptomic Correlates of Histology using Deep Learning. *bioRxiv*. Published online August 10, 2020:2020.08.07.241331. doi:10.1101/2020.08.07.241331

16. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*. Published online July 27, 2020:1-11. doi:10.1038/s43018-020-0087-6

17. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*. 2018;24(10):1559-1567. doi:10.1038/s41591-018-0177-5

18. Gutman DA, Cobb J, Somanna D, et al. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc*. 2013;20(6):1091-1098. doi:10.1136/amiajnl-2012-001469

19. 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*. Published online November 18, 2020:1-11. doi:10.1038/s41416-020-01122-x

20. Choi J-H, Hong S-E, Woo HG. Pan-cancer analysis of systematic batch effects on somatic sequence variations. *BMC Bioinformatics*. 2017;18. doi:10.1186/s12859-017-1627-7

21. Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*. 2017;35(6):498-507. doi:10.1016/j.tibtech.2017.02.012

22. Tom JA, Reeder J, Forrest WF, et al. Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics*. 2017;18(1):351. doi:10.1186/s12859-017-1756-z

23. Chatterjee S. Artefacts in histopathology. *Journal of Oral and Maxillofacial Pathology : JOMFP*. 2014;18(Suppl 1):S111. doi:10.4103/0973-029X.141346

24. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD. Removing Batch Effects From Histopathological Images for Enhanced Cancer Diagnosis. *IEEE J Biomed Health Inform*. 2014;18(3):765-772. doi:10.1109/JBHI.2013.2276766

25. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*. 2018;16:34-42. doi:10.1016/j.csbj.2018.01.001

26. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Computer Graphics and Applications*. 2001;21(5):34-41. doi:10.1109/38.946629

27. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE; 2009:1107-1110. doi:10.1109/ISBI.2009.5193250

28. Tellez D, Balkenhol M, Otte-Höller I, et al. Whole-Slide Mitosis Detection in H E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. *IEEE Transactions on Medical Imaging*. 2018;37(9):2126-2136. doi:10.1109/TMI.2018.2820199

29. Liu Y, Gadepalli K, Norouzi M, et al. Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv:170302442 [cs]*. Published online March 7, 2017. Accessed November 6, 2020. http://arxiv.org/abs/1703.02442

30. Tellez D, Litjens G, Bandi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*. 2019;58:101544. doi:10.1016/j.media.2019.101544

31. Anghel A, Stanisavljevic M, Andani S, et al. A High-Performance System for Robust Stain Normalization of Whole-Slide Images in Histopathology. *Front Med (Lausanne)*. 2019;6. doi:10.3389/fmed.2019.00193

32. Liu J, Lichtenberg T, Hoadley KA, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. 2018;173(2):400-416.e11. doi:10.1016/j.cell.2018.02.052

33. Thorsson V, Gibbs DL, Brown SD, et al. The Immune Landscape of Cancer. *Immunity*. 2018;48(4):812-830.e14. doi:10.1016/j.immuni.2018.03.023

34. Aggarwal N, K. Agrawal R. First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images. *JSIP*. 2012;03(02):146-153. doi:10.4236/jsip.2012.32019

35. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973;SMC-3(6):610-621. doi:10.1109/TSMC.1973.4309314

36. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:180203426 [cs, stat]*. Published online September 17, 2020. Accessed November 24, 2020. http://arxiv.org/abs/1802.03426

37. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:160204938 [cs, stat]*. Published online August 9, 2016. Accessed November 7, 2020. http://arxiv.org/abs/1602.04938

38. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv:181112808 [cs, stat]*. Published online December 2, 2018. Accessed August 19, 2020. http://arxiv.org/abs/1811.12808

39. Boyd S, Boyd SP, Vandenberghe L. *Convex Optimization*. Cambridge University Press; 2004.

40. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*. 2020;1(8):800-810. doi:10.1038/s43018-020-0085-8

41. Chen RJ, Lu MY, Wang J, et al. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *arXiv:191208937 [cs, q-bio]*. Published online September 3, 2020. Accessed November 24, 2020. http://arxiv.org/abs/1912.08937

42. Jaber MI, Song B, Taylor C, et al. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research*. 2020;22(1):12. doi:10.1186/s13058-020-1248-3

43. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *PNAS*. 2018;115(13):E2970-E2979. doi:10.1073/pnas.1717139115

44. Zadeh Shirazi A, Fornaciari E, Bagherian NS, Ebert LM, Koszyca B, Gomez GA. DeepSurvNet: deep survival convolutional network for brain cancer survival rate classification based on

histopathological images. *Med Biol Eng Comput*. 2020;58(5):1031-1045. doi:10.1007/s11517-020-02147-3

45. Wu Z, Wang L, Li C, et al. DeepLRHE: A Deep Convolutional Neural Network Framework to Evaluate the Risk of Lung Cancer Recurrence and Metastasis From Histopathology Images. *Front Genet*. 2020;11. doi:10.3389/fgene.2020.00768

46. Wulczyn E, Steiner DF, Xu Z, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS ONE*. 2020;15(6):e0233678. doi:10.1371/journal.pone.0233678

47. Luo X, Zang X, Yang L, et al. Comprehensive Computational Pathological Image Analysis Predicts Lung Cancer Prognosis. *Journal of Thoracic Oncology*. 2017;12(3):501-509. doi:10.1016/j.jtho.2016.10.017

48. Jiang D, Liao J, Duan H, et al. A machine learning-based prognostic predictor for stage III colon cancer. *Scientific Reports*. 2020;10(1):10333. doi:10.1038/s41598-020-67178-0

49. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*. 2019;16(1):e1002730. doi:10.1371/journal.pmed.1002730

50. Courtiol P, Maussion C, Moarii M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat Med*. 2019;25(10):1519-1525. doi:10.1038/s41591-019-0583-3

51. Daly B, Olopade OI. A perfect storm: How tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. *CA: A Cancer Journal for Clinicians*. 2015;65(3):221-238. doi:10.3322/caac.21271

52. Bickell NA, Wang JJ, Oluwole S, et al. Missed Opportunities: Racial Disparities in Adjuvant Breast Cancer Treatment. *JCO*. 2006;24(9):1357-1362. doi:10.1200/JCO.2005.04.5799

53. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med*. 2018;378(11):981-983. doi:10.1056/NEJMp1714229

54. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*. 2015;10(3). doi:10.1371/journal.pone.0118432

55. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

56. Magee D, Treanor D, Crellin D, et al. Colour Normalisation in Digital Histopathology Images. *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*. Published online January 1, 2009.

57. Tang TT, Zawaski JA, Francis KN, Qutub AA, Gaber MW. Image-based Classification of Tumor Type and Growth Rate using Machine Learning: a preclinical study. *Scientific Reports*. 2019;9(1):12529. doi:10.1038/s41598-019-48738-5

58. Zanjani FG, Zinger S, Bejnordi BE, Laak JAWM van der, With PHN de. Stain normalization of histopathology images using generative adversarial networks. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. ; 2018:573-577. doi:10.1109/ISBI.2018.8363641

59. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

60. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337. doi:10.1038/nature11252

61. Hammerman PS, Lawrence MS, Voet D, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525. doi:10.1038/nature11404

62. Collisson EA, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-550. doi:10.1038/nature13385

63. Creighton CJ, Morgan M, Gunaratne PH, et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*. 2013;499(7456):43-49. doi:10.1038/nature12222

64. Lawrence MS, Sougnez C, Lichtenstein L, et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576-582. doi:10.1038/nature14129

65. Carrot-Zhang J, Chambwe N, Damrauer JS, et al. Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell*. 2020;37(5):639-654.e6. doi:10.1016/j.ccell.2020.04.012

66. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv:161002357 [cs]*. Published online April 4, 2017. Accessed August 12, 2020. http://arxiv.org/abs/1610.02357

67. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2009:248-255. doi:10.1109/CVPR.2009.5206848

68. Gonzalez RC, Woods RE. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc.; 2006.

69. IBM. *IBM ILOG CPLEX 12.10 User's Manual*. IBM ILOG CPLEX Division, Incline Village, NV; 2017.