Impact of Stain Normalization on Computer-Aided Survival Prognosis using HE-Stained Whole-Slide Images in Breast Cancer

Chris-Andris Görner[1]

[1]Medical Informatics, Heilbronn University of Applied Sciences & University of Heidelberg, Germany.

Abstract

Variations in hematoxylin and eosin (H&E) staining pose a challenge for deep learning models in computational pathology. While stain normalization is widely studied for computer-aided diagnosis (CAD), its impact on computer-aided prognosis (CAP), particularly survival prediction, remains underexplored. This study investigates how different stain normalization and augmentation techniques affect feature extractor-based survival prognosis models using whole-slide images (WSIs) in breast cancer. We reformulate the CLAM architecture for survival prognosis, integrating both WSIs and clinical data from the TCGA-BRCA dataset. Five commonly used stain normalization and augmentation methods, including Reinhard, Macenko, Tellez's augmentation, StainGAN, and MultiStain-CycleGAN, were evaluated. Performance was assessed with a 10-fold cross-validation using the concordance index (C-index), comparing models trained on clinical data alone and with unnormalized and normalized/augmented WSI features. Stain normalization exhibited mixed effects. Common techniques like Reinhard and Macenko did not improve model performance, consistent with earlier studies. Stain augmentation and the MultiStain-CycleGAN yielded modest performance gains using only preclinical features. However, adding histopathological features to models with strong diagnostic clinical data did not provide additional benefit, indicating that well-established clinical markers remain dominant predictors. Our findings suggest that conventional stain normalization adds limited value for survival prediction when combined with robust clinical features and modern feature extractors.

Keywords: stain normalization, computer-aided prognosis, survival prediction

1 Introduction

Histopathological analysis represents a crucial step in the diagnosis of cancerous diseases through the examination of tissue samples. Expert pathologists derive key aspects of a disease by inspecting specially prepared tissue sections under a microscope, including tumor malignancy, cancer stage, molecular profiling and more. Since the advent of whole-slide imaging (WSI) through the introduction of the first commercially available whole-slide scanners around the turn of the millenium, i.e. high resolution scanning of prepared tissue slides, as well as the advent of deep learning (DL), computer vision in histopathology has been of increasing interest as a decision support tool, commonly called computational pathology (CPATH). [1] Current applications include tumor detection, segmentation, classification and grading as well as cell detection and counting among others. [2]

The preparation and scanning of tissue slides is a complex task. One crucial step of slide preparation is staining, which increases contrast and highlights cell structures of interest. Although a wide variety of stains exist, it is estimated that around 80% of all stains are Hematoxylin and Eosin (H&E or HE) stains, imparting the tissue with a pinkish-blue color. [3] Depending on the staining protocol and scanner used, stain color can vary drastically. These variations pose unique challenges to CPATH systems. Model generalizability and robustness is a general struggle of DL models, however CPATH is especially prone to internalize biases induced by artifacts such as stain color variations. [2, 4] Techniques to mitigate the variation of stain color post digitization have been a subject of active research. The purpose of stain normalization or augmentation is to eliminate color variations, driving CPATH model gradients by image content and not color variations or artifacts and accordingly promoting robust and generalizable models. Although several meta-analyses have been conducted on the impact of stain normalization on the performance of computer-aided diagnosis, there is no clear consensus on the merits of stain normalization among the scientific community. Additionally, continuous advances in DL research and new state-of-the-art stain normalization techniques warrant frequent comparison. Therefore, we find it adequate to take a closer look at recent stain normalization and augmentation techniques for HE stains.

Apart from computer-aided diagnosis, prognostic models focus on predicting a patient's risk to experience events such as disease progression or ultimately survival. Katzman et al.'s DeepSurv [5] demonstrated in 2016 that DL-based survival prognosis models can outperform traditional statistics-based models on clinical data, adapting the traditional linear Cox proportional hazards model to serve as a loss function for the network. Swiftly after, Zhu et al.'s DeepConvSurv [6] extended this work by directly inputting WSI patches into a Convolutional Neural Network (CNN) and outputting a risk score. Today, prognostic models are typically multi-modal, capturing the relationship between diagnostic slides, clinical and omics data as morphological correlates. It has been shown that DL-driven patient stratification into risk groups can lead to better separation than traditional discriminative biomarkers such as tumor grade or stage. The risk sub-cohorts offer opportunities to clinicians for prioritization and tailored therapies. [7] Computer-aided prognosis is an emerging field that promises high

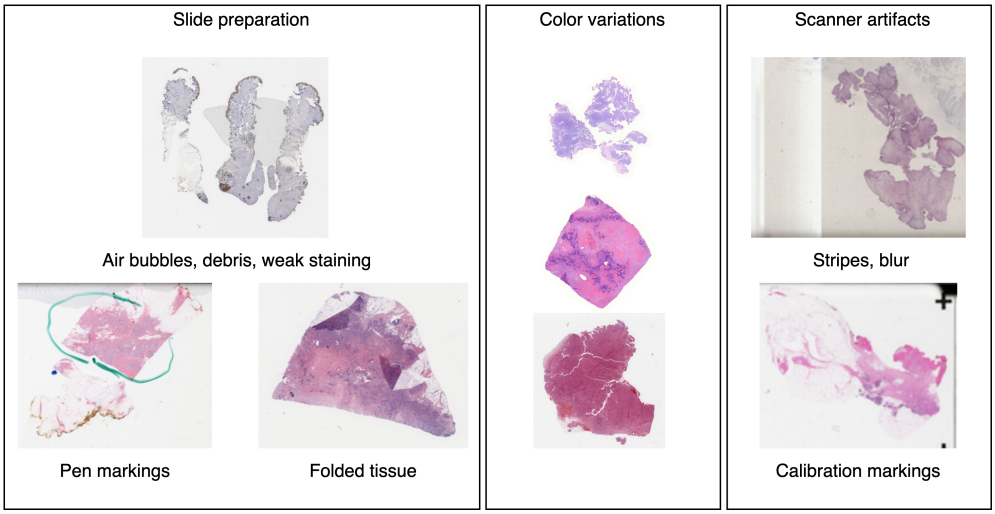

Fig. 1 Artifacts present in the WSIs of the TCGA-BRCA dataset.

clinical applicability. On the flipside, to the best of our knowledge it has not been studied how stain normalization can affect prognostic accuracy.

This work aims to contribute to the elucidation of the effect of stain normalization on the performance of deep learning-based prognostic models. We evaluate several stain normalization methods on the TCGA-BRCA dataset in a survival prediction task using deep learning. The TCGA-BRCA dataset is chosen due to its public availability of roughly one thousand WSIs of breast cancer patients of from 42 tissue source sites including clinical data and patient survival annotations.

2 Background

2.1 WSI preparation & artifacts

Several steps are needed to produce long-lasting diagnostic slides. First, tissue undergoes fixation, which impedes degradation and retains the cell structure of the tissue. Fixatives achieve this by irreversibly cross-linking proteins. To facilitate the cutting of thin sections, the fixed tissue is then dehydrated. Embedding further enhances the structural integrity for the subsequent sectioning. In case more advanced immunohistochemical (IHC) stains are employed in order to identify the presence of certain antigens, antigen binding sites previously masked by fixatives have to additionally be retrieved. [8] Finally, staining is employed to highlight certain structures of the tissue and increase contrast to facilitate visual analysis. Hematoxylin is a basic dye that stains acidic structures, imparting cells and nuclei with a purplish or bluish color. Eosin is an acidic dye that is subsequently applied to stain acidophilic structures, mainly muscle fibers, cytoplasm and collagen. [8, 9]

A wide range of tissue preparation inconsistencies can impede successful application of computational pathology methods, occurring during tissue biopsy, fixation and staining, visualized in Fig. 1. HE staining is particularly susceptible to color inconsistencies determined by dye concentration and application technique across institutions. Moreover, the choice of scanner, its calibration, the chosen scanning protocol and interfering artifacts such as dust and air bubbles can significantly alter the appearance of the WSI. [10, 11] HE color is generally thought to have a high influence on CPATH systems, since it not only increases contrast but directly binds to specific cell structures, carrying morphological information. Thus, simply converting the RGB image to grayscale would significantly reduce the tumor signal. [12] Consequently, stain normalization techniques have been and continue to be of scientific interest.

2.2 Categories of stain normalization techniques

Stain normalization approaches can generally be distinguished into three categories, as described by Hoque et al. [9].

Histogram transformation-based approaches employ statistical matching of the color space by aligning the statistical distributions of the source image to a target color distribution. The reference color space is usually tuned to represent a sort of balanced or mean stain representation of a large collection of WSIs. As Ke et al. underlines, a major drawback of this is that the normalization performance depends heavily on the chosen reference color space. [13] These techniques are relatively computationally efficient, however, because they treat the source image as a whole without explicitly separating the contribution of different stains, they can introduce artifacts and may distort biologically relevant structures, especially when the source and reference color distribution differ substantially.

Deconvolution-based approaches operate by decomposing an image into its underlying stain components, typically using an optical density model that represents the image as a linear combination of a stain and a concentration vector. These methods can further be distinguished as separated and unified: separated methods estimate and normalize each stain channel independently, while unified methods apply the transformation jointly across all channels in a single step, often within a latent space. These methods can be further categorized based on whether the stain matrix is known a priori or learned directly from the data, with techniques such as non-negative matrix factorization enabling unsupervised decomposition. These approaches generally improves stain-specific consistency. Nonetheless, the process involves multiple transformation steps, which can be computationally expensive and susceptible to cumulative error. Additionally, if one stain is less pronounced or missing, the separation can be unreliable, leading to poor normalization quality.

Generative adversarial network-based approaches leverage generative adversarial networks (GAN) where color normalization is treated as an image-to-image translation task. GAN-based models learn the staining by internalizing the mapping to a target domain so that the adversarial discriminator deems the generated image indistinguishable. They are capable of capturing complex, nonlinear relationships between staining appearances while maintaining tissue morphology. Nevertheless, they require

large and diverse datasets for stable training, are sensitive to hyperparameter tuning, and can sometimes produce unrealistic colors or lose fine-grained structural details.

2.3 Recent literature on stain normalization impact

For the review of current stain normalization techniques, we majorly lean on the 2023 systematic review by Hoque et al. [9]. They review 35 stain normalization techniques and select 10 for further analysis beyond suggesting their own approach. Hoque et al. concludes that no general suggestion of a specific type of stain normalization can be made. According to their findings, some techniques even introduce new kinds of artifacts. They favor deconvolution-based approaches, which their proposed approach can also be categorized as. In spite of their detailed analyses, they assess the effects of the selected techniques solely based on similarity measures such as normalized median intensity (NMI) and the structure similarity index metric (SSIM) without consulting CAD or CAP tasks. Although some correlation between similarity measures and CPATH performance can be expected, a strong relationship between the two without empirical validation on downstream tasks cannot be assumed.

To this end, recent works that investigate CPATH performance in light of different stain normalization techniques include Tellez et al. [14], Voon et al. [15] and Boschman et al. [16].

Tellez et al. examines both stain color normalization as well as augmentation on relevant clinical tasks, namely mitotic figure detection, tumor metastasis detection, prostate epithelium detection and colorectal cancer tissue type classification. They deduce color normalization to be negligible in favor of a combination of basic, morphological and color deconvolution-based augmentation.

Voon et al. concentrates on invasive ductal carcinoma grading, evaluating six stain normalization techniques, namely Reinhard [17], Macenko [18], Vahadane [19], Adaptive Color Deconvolution (ACD) [20], StainGAN [21] and StainNet [22], finding consistently worse performance with respect to an unnormalized baseline. They employ several state-of-the-art CNNs as feature extractor with a dense trainable classification head. They suggest stain normalization to not be as beneficial to CPATH performance as previously considered.

Boschman et al. investigates eight stain normalization techniques on tumor detection and grading, namely Reinhard, Macenko, Vahadane as well as Grayscale Vahadane, ACD, Histspec [23], Khan [24] and Zanjani [25], distinguishing between performance on held-out test data from the same tissue source site as well as an external site. They also suggest a new approach by effectively combining the Reinhard, Vahadane and Macenko techniques. They confirm no consistent performance increase when evaluating on images from the same site that the classifiers were trained on, however they do find a consistent improvement when evaluating on external images. Moreover, they find that their combined normalization approach enhances generalization.

A thorough evaluation of how stain normalization techniques impact CPATH outcomes is essential to draw meaningful conclusions about their practical utility. Seeing

as the findings of these works are conflicting, this review concludes that further investigations are warranted. Additionally, none of the mentioned impact studies consider computer-aided prognosis, which this work will contribute to.

2.4 Review: Applications of stain normalization in CAP

A small literature review on stain normalization techniques in recent survival prognosis studies is also conducted. The focus of this review is to estimate to which extent stain normalization is employed, whether its contribution to model performance is assessed and what these assessments conclude. The flowchart in Appendix A Fig. A1 illustrates the review process following the PRISMA 2020 guideline for reporting systematic reviews [26]. Articles were extracted from the Scopus, Pubmed and IEEE Xplore database using the search string in Listing 1. From 134 initially identified articles, 6 were ultimately included.

Three of the included studies utilized stain normalization or augmentation in the preprocessing of WSIs, however they did not conduct ablation studies to assess the effect on model performance. Two of the studies employed Macenko stain normalization [27, 28], while the other adopts Vahadane stain normalization [29].

Among the three studies that did conduct ablation studies, two noted a positive effect of employing stain normalization and one remarked worse performance over an unnormalized baseline. Elforaici et al. propose their own GAN-based stain style transfer normalization for HE and Hematoxylin Phloxine Saffron stained slides for a semi-supervised vision transformer knowledge distillation network, registering a consistent performance boost when employing their normalization approach. [30] Ma et al. applies a variant of Macenko normalization with subsequent basic data augmentation, performing survival prognosis with a binned random forest using features extracted by a CNN. [31] Jiao et al. on the other hand notice that Macenko stain normalization can introduce artifacts and degrade model performance in their pipeline, performing tissue segmentation to extract tumor microenvironment features used for survival analysis. [32]

Among the studies reviewed, the majority appear to select a normalization technique without providing a clear rationale or conducting ablation studies, often defaulting to one of the three most commonly used methods (Reinhard, Macenko, Vahadane). However, recent comparative analyses have demonstrated that these conventional techniques do not consistently improve model performance.

2.5 Methods selected for investigation

The methods selected for investigation are shown in Table 1. Selection criteria include a readily available code implementation with a moderate runtime, preferably with GPU support. For GAN-based methods, models with pre-trained weights for HE stain normalization are of interest. Reinhard and Macenko were chosen for their popularity. Given the meta-analysis by Tellez et al., their recommended method is also included. For GAN-based approaches, Shaban et al.'s StainGAN with pre-trained weights on the CAMELYON16 challenge [33] for lymph node metastasis detection is included as they were the first to leverage the CycleGAN architecture for stain normalization.

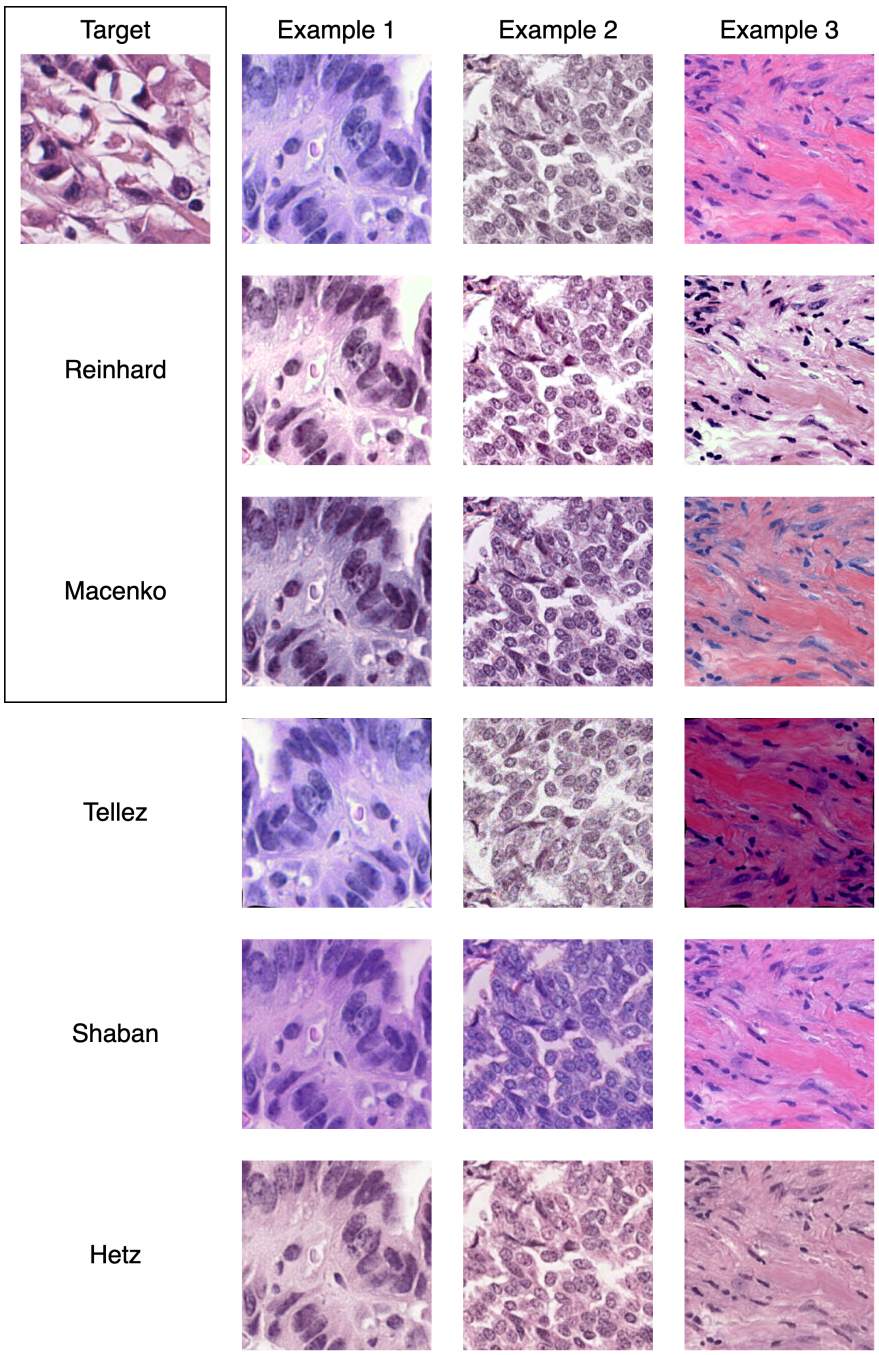

Fig. 2 Overview of selected stain normalization techniques. Reinhard and Macenko stain normalization require a target domain, i.e. a target patch in this case.

Following a search of relevant databases, Hetz et al.'s proposes another GAN-based approach that is included for recency, pre-trained on the CAMELYON17 challenge [34] for the detection and classification of breast cancer metastases. Fig. 2 depicts the application of each of these methods on three example patches.

Paper	**Given name**	**Publication Year**	**Architecture**	**Code availability**
Reinhard [17]	-	2001	Statistical matching	TIAtoolbox
Macenko [18]	-	2009	Color deconvolution	TIAToolbox
Tellez [14]	-	2019	Stain augmentation	no, but replicable from paper
Shaban [21]	StainGAN	2019	CycleGAN	GitHub[1]
Hetz [35]	MultiStain-CycleGAN	2024	CycleGAN	GitHub[2]

[1] https://github.com/xtarx/StainGAN

[2] https://github.com/DBO-DKFZ/multistain_cyclegan_normalization

Table 1 Stain normalization/augmentation methods selected for comparison.

Reinhard et al. stain normalization is not specific to HE staining as was originally conceptualized for any type of color transfer. The color distribution of an image is matched that of a target image by aligning their means and standard deviations in the Lab color space. It relies on the assumption that the variations in histopathological staining can be captured and corrected using global color statistics.

Macenko et al. stain normalization is based on the Beer-Lambert law, which describes the relationship between the absorption of light as proportional to the concentration of the substance [36], assuming that each pixel is a linear combination of stain contributions. It estimates stain vectors through singular value decomposition (SVD) in the optical density (OD) space and performs color deconvolution and reconstruction to match stain concentrations with a target.

Tellez et al. stain augmentation is the result of their meta-analysis that concluded with stain normalization being negligible and stain augmentation to be favorable in most scenarios. Alongside traditional data augmentation techniques such as basic and morphological transformations, they find random perturbations of the H and E channels in Haematoxylin-Eosin-DAB (HED) color space, again obtained by transforming the RGB image into OD space and applying color deconvolution using a known stain matrix, to be most beneficial to model robustness.

Shaban et al.'s StainGAN leverages the CycleGAN architecture to perform unsupervised image-to-image translation between stain domains. Two generators convert between stain domains and two discriminators distinguish real images in the target stain domain with generated ones. Adversarial loss makes sure that generated images look like those of the target domain. Cycle consistency loss ensures images look the same after conversion to the target domain and back to the source domain, enabling the

preservation of tissue structures. Lastly, identity loss guarantees that images already in the target style remain unchanged.

Hetz et al.'s MultiStain-CycleGAN similarly employs a CycleGAN architecture with some key differences, enabling images from different source domains to effectively be converted to a target domain without retraining. They convert input images to 3-channel grayscale and apply strong color augmentations to simulate a wide variety of stain appearances, enabling domain-invariant normalization. They introduce an intermediate domain to help map all input domains into a common space, simplifying the mapping to the target domain and replacing the identity loss with a reconstruction loss from grayscale input to improve learning from augmented data.

3 Research Gap

The recent reviews and impact studies presented in Section 2.3 provide a strong foundation for this work. Out of these studies, none assessed the impact on CAP with only Voon et al. evaluating CNN-based feature extractors at all to the best of our knowledge. Their increasing popularity warrant further analysis. Moreover, when [stepping into] the domain of survival prognosis, it is often not of interest to perform prognosis based on imaging data alone but to include medical imaging as part of a wide range of multi-modal inputs. We want to additionally model the clinical workflow and gauge the interaction between clinical and image data in light of stain normalization. To this extent, the research question can be asked:

What effect do stain normalization techniques have on the performance of CNN-based feature extractor driven survival prognosis from whole slide images in unison with clinical data?

4 Methods

4.1 Data acquisition

Whole-image slides were obtained from the The Cancer Genome Atlas Program (TCGA) Breast Invasive Carcinoma (BRCA) project through the Genomic Data Commons (GDC) Data Portal.[3] Corresponding clinical data was retrieved from cBioPortal. The Firehose Legacy study was selected since it consists of the most exhaustive clinical information for the TCGA-BRCA project.
TCGA-BRCA comprises 1098 diagnostic slides. Upon visual examination, the majority of slides from the tissue source sites A1, B6 and E9 exhibit marker annotations as shown in Fig. 1, which is a considerable artifact, thus they are excluded. Further excluding slides without survival annotations, 924 WSIs from 36 different tissue source sites remain. Among the 924 patients analyzed, 110 deaths were observed during the follow-up period, with the remaining 814 patients being right-censored.

[3] https://portal.gdc.cancer.gov/

An examination of the distribution of key clinical markers of the cohort (see Appendix B Fig. B2) sheds light on several biases. Due to the nature of the disease, male patients are severely underrepresented, possibly leading to more inaccurate prognostic outcomes. The distribution of patient race is severely unequal, with white patients making up more than two thirds of the entire cohort. It has been shown that patient race and the tissue site can be inferred from histopathological images even after applying Reinhard, Macenko or Vahadane stain normalization. [37] The need for more balanced survival studies with larger cohorts will be touched on in the discussion in chapter 6.

4.2 Pipeline

The prognostic pipeline is constructed with the aim of providing a simple but capable model based on widely accepted tools and architectures in order to facilitate a high level of comparability with other studies. The primary goal of the pipeline is to minimize the impact of the specific choice of model architecture on performance and maximize the dependency of results on the normalization techniques. For this reason, an adaptation of Clustering-constrained Attention Multiple Instance Learning (CLAM) [38] for survival analysis is employed, see Fig. 3. CLAM is a weakly supervised architecture for slide-level predictions based on slide-level annotations without the need for patch- or ROI-level annotations. It leverages an attention-based multiple instance learning architecture which enables the model to identify regions of pathological interest itself. Possibly the most debilitating challenge of using diagnostic slides as input for deep learning models are their size, with WSIs being up to 100,000 times larger than ImageNet samples measuring 256x256 pixels, rendering the handling of WSIs on current GPUs difficult. The loss of considerable detail by either downsampling or choosing a different magnification level is counter-productive. Therefore, it is common practice to divide WSIs into smaller patches, consequently shifting the problem to obtaining WSI-level predictions from patch-level inputs. [7] In its original release, CLAM was designed for classification tasks, however it can be adapted to perform survival prognosis by integrating a loss function adapted to time-to-event data. For comparison with their own model, Yang et al. [39] recently published a modified CLAM repository[4] for survival prognosis, which this work leans on.

4.2.1 Patch creation

For the segmentation of the foreground tissue and the creation of the patches, the CLAM pipeline utilizes the OpenCV[5] library. Each digitalized slide is loaded at a downsampled level and converted from RGB to HSV color space. Median blurring, thresholding and morphological closing are applied before the contours are filtered and patches are created by tiling the detected foreground tissue into squares of a specified size. In preparation for feature extraction, the coordinates of the identified patches are saved. The process can be tuned by several hyperparameters. CLAM provides a preset configuration adapted to TCGA, thus it is adopted in this work.

[4] https://github.com/Zhcyoung/BEPH
[5] https://opencv.org/

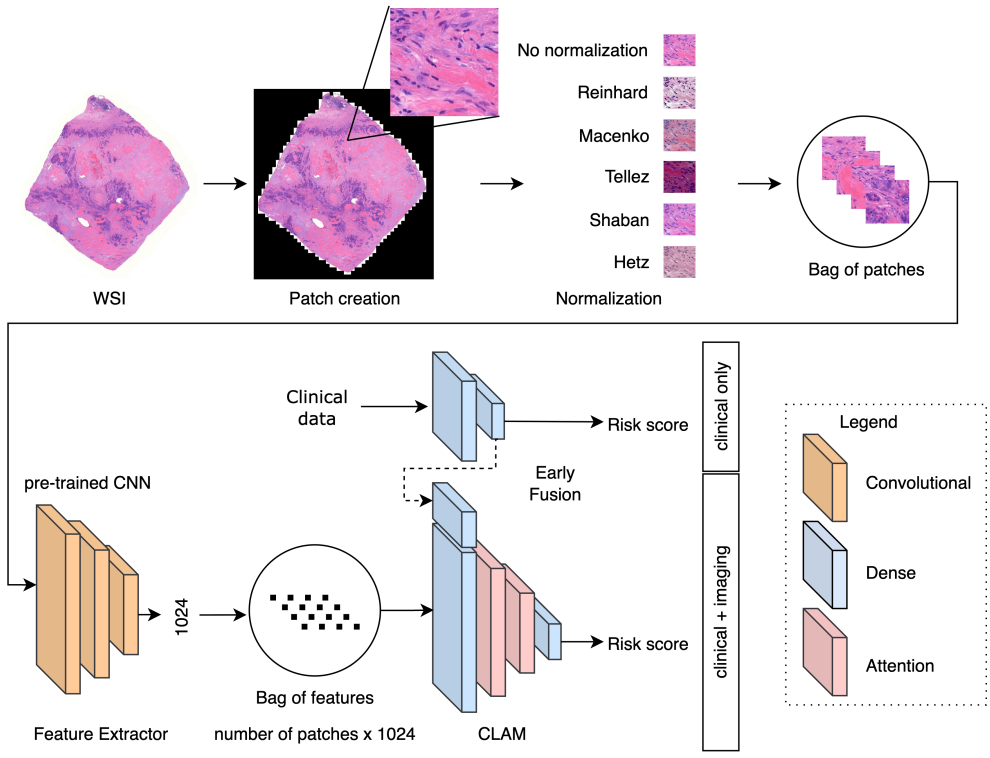

Fig. 3 CLAM pipeline for survival prognosis.

4.2.2 Feature extraction

Patches are loaded in batches at full resolution from the saved coordinates. Then, the selected stain normalization techniques are applied. For Reinhard and Macenko stain normalization, a target patch is selected from the WSI which exhibits a color distribution closest to the mean of all WSI colors in LAB space. A low-dimensional feature vector is then extracted by passing each patch through a pre-trained computer vision backbone. The choice of feature extractor is a considerable decision, although backbones of histopathological foundation models, often based on vision transformers, promise especially information-rich feature extractions, they come at the expense of a much greater computational cost. Accordingly, the ResNet50 CNN was elected as feature extractor for this work. Namely, two-dimensional adaptive average pooling is applied after the third residual block of the ResNet50 backbone pre-trained on ImageNet, resulting in a 1024 feature vector per patch. Normalization is applied on the fly to each patch before passing it to the extractor backbone. The dimensionality reduction of the input enables loading the patch-wise feature representation of an entire slide into the memory of a commercially available GPU, effectively circumventing the need for restrictive patch aggregation or sampling.

4.2.3 Clinical data

Since the goal of this work is to gauge the impact of stain normalization on histopathological images in survival prognosis, the baseline model will only be trained and evaluated using features extracted from images. However, the contribution of tabular clinical data should also be considered, especially concerning its rich information content harnessed in multi-modal models.
To reduce noise, only clinical data with less than 10% missing values is considered. Missing values are imputed by employing a multivariate imputer, categorical features are one-hot encoded and numeric features are normalized to the range $[0, 1]$ using a min-max scaler to ensure ensure the comparability of feature scales and to facilitate the stability of model convergence during training. Consulting the clinical data available from CBioPortal, the selected features are listed in Table 2.

Feature	**Data Type**	**Value Range**
Age at Diagnosis	int	33–62
Sex	categorical	Female, Male
Prior Cancer Diagnosis Occurrence	categorical	Yes, No
Menopause Status	categorical	Pre, Peri, Post, Indeterminate
Race Category	categorical	White, Black or African American, Asian, American Indian or Alaska Native
*Estrogen Receptor Status by IHC	categorical	Positive, Negative, Indeterminate
*Progesterone Receptor Status by IHC	categorical	Positive, Negative, Indeterminate
*Fraction of Genome Altered	float	0–1
*AJCC Tumor Stage	categorical	T1, T2, T3, T4
*AJCC Lymph Node Stage	categorical	NX, N0, N1, N2, N3
*AJCC Metastasis Stage	categorical	MX, M0, M1

Table 2 TCGA-BRCA clinical features made available by CBioPortal included. Features marked with * are considered diagnostic clinical data.

The fine-grained diagnostic classifications in American Joint Committee on Cancer (AJCC) stage codes are mapped to the overall stage, e.g. N0 (i-) to N0, to encourage generalization. Additionally, the one-hot encoded menopause stage is set to zero for male patients.
To further assess the impact of including clinical data in the pipeline, a distinction will be made between data collected at admission or through patient self-report and data collected through targeted laboratory testing or clinical evaluation by a physician, hereinafter referred to as preliminary clinical data and diagnostic clinical data respectively. The latter is marked with * in Table 2. This distinction reflects real-world clinical workflows, where initial decision-making is often guided by readily available

information, while subsequent prognostic assessments may incorporate more targeted diagnostic findings and imaging results.

4.2.4 Model architecture

The CLAM model is empowered by several deep learning paradigms. Multiple instance learning is essential to leveraging the weakly supervised WSI-level labels using patch-wise instances. Instead of each training sample being labeled individually as it is the case in strongly supervised learning, in MIL a bag of instances with a bag-level label, i.e. a collection of patches with a slide-level label in this case are provided. Patch-level predictions are aggregated to a slide-level prediction on which the loss is calculated. An attention mechanism learns the contribution of the patches towards the slide-level risk. Clustering-constrained instance evaluation is a powerful tool to encourage an even more discriminative learning by inferring patch labels from the attention mechanism and incorporating a direct patch-level loss. In our experiments, instance evaluation lead to consistently inferior performance on the unseen test set compared to purely relying on the bag-wise loss, therefore it is disabled in this work.

The initial CLAM release was conceptualized specifically for inputs from feature extractors. In the interest of introducing clinical data into the model, several architectural options can be considered. For simplicity, the attention mechanism processes the image-related features as usual, while a separate simple neural network with two hidden layers encodes representations of the clinical features. In a process commonly called early fusion, the logit output of the last hidden layer of the dense clinical net is concatenated with the feature representation of all slide patches. Then, the clinical and WSI features are being passed through the attention mechanism, enabling the model to internalize important patches in unison with clinical data.

4.2.5 Loss function

Cox Proportional Hazards (Cox PH) or Cox regression formalized in 1972 by statistician David Cox is a multivariate statistical technique which models the hazard function of at-risk individuals as a regression problem[40]. The hazard function is a function of time $h(t)$ of explanatory variables or covariates x_i and regression coefficients β_i multiplied by a baseline hazard $h_0(t)$ for all individuals i in cohort N, see Eq. 1.

$$h(t) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i) \tag{1}$$

The baseline hazard $h_0(t)$ is an arbitrary function that represents the overall hazard if all coefficients are zero. Cox PH is referred to as semi-parametric, as it combines parametric components of the covariates with the non-parametric baseline hazard. The hazard of the event at any point in time is a multiple of the baseline hazard. This in turn signifies that the hazards of each individual are proportional to each other and cannot intersect over time. The proportional hazards assumption is the main limiting factor of Cox regression as it cannot be assumed that hazard ratios never change over time, especially in a clinical setting given treatments, surgical interventions, etc. Nonetheless, Cox regression provides a valuable approximation to a multi-faceted problem. [41]

Supposing that $h_0(t)$ is arbitrary, Cox proposes the partial likelihood in order to estimate the coefficients β of the risk function h, see Eq. 2. Each individual i has an event indicator $\delta_i \in \{0, 1\}$ (1 = event occurred, 0 = censored), covariates X_i and risk set $R(T_i)$ comprising all individuals still at risk at time T_i.

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{exp\left(\beta X_i\right)}{\sum_{j \in R(X_i)} exp\left(\beta X_j\right)} \right]^{\delta_i} \tag{2}$$

In the context of formulating Cox regression for deep learning, the linear predictor βX is replaced by a neural network output as formalized by Faraggi and Simon et al.[42] given the model parameters θ and the estimated risk function $\hat{h}_\theta$ by the deep learning model. Taking the logarithm of the partial likelihood is advantageous since it to converts the product into a sum, providing better numerical stability and enhancing convexity, which is desired for optimization tasks and simplifies gradient computation in machine learning applications. To phrase the log likelihood as a minimization problem, a negative sign is placed in front. Consequently, the negative log partial likelihood is employed as a loss function for the CLAM model, see Eq. 3.

$$l(\theta) = -log\left(L(\theta)\right) = -\sum_{i=1}^{n} \delta_i \left(\hat{h}_\theta(x_i) - log \sum_{j \in R(x_i)} exp\left(\hat{h}_\theta(x_j)\right) \right) \tag{3}$$

The negative log partial Cox likelihood loss encourages the model to give higher risk scores to individuals x_i who experience the event earlier than the remaining at-risk individuals x_j.

4.2.6 Training & Validation

The ratio imbalance of censor-to-event data as well as the need for site preservation across splits to accurately assess the success of stain normalization poses a significant challenge. Stratification of a feature by preserving the feature distribution in all splits is a straightforward task. However, ensuring that WSIs from every tissue source site are preserved within either the training or the test set in combination with retaining an approximate feature distribution is a challenging problem, seeing as any given tissue source site in the TCGA project submits cohort data of highly variable size and censor-to-event ratios. Howard et al. [37] formalizes this problem in their work and provides a convex optimization/quadratic programming approach to create k-folds with preserved sites and stratification by a desired feature, which we adopt here. It is important to note that, due to the nature of the problem, the sizes of the splits only approximately adhere to the desired proportions. Using their method, a 5-fold cross validation 70%/10%/20% train/validation/test split is constructed. The model is trained for a minimum of 20 epochs with early stopping triggering after 5 epochs of no increase in the validation score. All further hyperparameters are set to their default value.

Concerning the training methodology, the original CLAM model applies bag-wise gradients. To address training instability and poor convergence behavior observed during initial experiments - characterized by the model oscillating around a concordance index of approximately 0.5 across epochs - we employed gradient accumulation. This technique enables meaningful training with a larger effective batch size without violating memory constraints. It must be noted that due to the extremely variable bag size, the number of accumulation steps that can be sustained is highly dependent on the available RAM. We find that accumulating gradients over 10 steps provides the best trade-off between stability and performance given our setup, though this choice is based primarily on empirical observation and is subject to optimization in future considerations.

4.2.7 Evaluation

For evaluation, the concordance index (C-index) is a characteristic metric in time-to-event analysis, see Eq. 4. It measures how often on average a model predicts two distinct individuals to exhibit an ordering of risks that reflect the observed event time. Generally, the concordance index is calculated as the proportion of concordant pairs, i.e. pairs of individuals that are predicted to experience the event in a certain order, among all admissable pairs, i.e. pairs of individuals actually experiencing the event in a certain order. All admissable pairs are made up of distinct individuals i and j of cohort N for which i experiences the event before j either also experiences the event or is censored. Concordant pairs are a subset of admissable pairs where the risk ordering $\eta_i > \eta_j$ reflects observed event times $T_i < T_j$. $\mathbb{I}(\cdot)$ is the indicator function, evaluating to 1 if the condition is met, otherwise 0. Several versions of the concordance index exist with slight variations to accommodate shortcomings of the metric proposed by Harrell et al. in 1982 [43]. Since the original formula is sensitive to overestimating the performance on highly censored data, this work utilizes an extension by Uno et al. [44] that is generally more robust. Uno et al. incorporates a so-called inverse probability of censoring weights (IPCW) by weighting each pair by the Kaplan-Meier estimate $\hat{G}(\cdot)$ of the censoring survival function of the earlier subject's event time T_i. Uno's C-index dampens the contribution of pairs in high censoring regions since they can't confidently be assessed.[45] To ensure a well-calibrated IPCW, it is common to use the censoring distribution of the training data for the Kaplan-Meier estimate on the independent test set.

$$\widehat{C}_{\text{Uno}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \cdot \frac{\mathbb{I}(T_i<T_j)\cdot\mathbb{I}(\eta_i>\eta_j)}{\hat{G}(T_i)^2}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \cdot \frac{\mathbb{I}(T_i<T_j)}{\hat{G}(T_i)^2}} \tag{4}$$

The concordance index consolidates censoring distributions by considering pairs of uncensored and censored individuals where the event occurred before censoring. Only pairs of two uncensored individuals are excluded since no realistic assumptions about their risk can be concluded. The concordance index ranges from 0 to 1, with 1 indicating perfect ordering and 0 expressing complete anti-concordance or perfect reverse ordering. A concordance index of 0.5 reflects random ordering.

To test for statistical significance, the Wilcoxon signed-rank test is employed. It is a non-parametric test for paired data where a normal distribution cannot be assumed, which is adequate to compare the means of k-fold cross validations. The null hypothesis will be rejected for $p < 0.05$. The hypotheses to be tested are as follows:

H0: The inclusion of histopathological imaging features with or without stain normalization does not improve model performance over the baseline trained on tabular clinical data only. $\mu \leq \mu_0$

H1: $\mu > \mu_0$

5 Results

The results are subsequently presented and distinguished by training on preliminary or diagnostic clinical data, as described in Section 4.2.3. In each case, the baseline models are only trained and evaluated on tabular clinical data. The baselines are compared with the inclusion of histopathological imaging without and with the selected stain normalization techniques in the WSI feature extraction preprocessing. The results of the 10-fold cross validation are visualized in Fig. 4 and listed in Table 3.

First, both the preliminary and diagnostic clinical data baselines do not consistently achieve a concordance index of 0.5 on the held-out test set. This underlines the stark contrasts among the study cohorts between tissue source sites as well as the model's struggle to generalize from the seen training data, with or without training on imaging features. This goes to show that the model does internalize biases present in the dataset and, that the available data is most likely not sufficient to capture all elements related to patient survival. With the inclusion of imaging features in the training, the attention mechanism can extract some complementary signal from WSIs in the preclinical context, which is especially amplified by Tellez et al.'s stain color augmentation (0.635 ± 0.086) as well as Hetz et al.'s MultiStain-CycleGAN (0.652 ± 0.069). Interestingly, replacing the generic ResNet backbone with the histopathology-specific UNI model [46] as feature extractor did not yield improvements in this survival prediction task. This suggests that domain-specific pretraining alone is not sufficient when the pretraining task is poorly aligned with survival prognosis, highlighting the importance of task-specific finetuning and flexible feature adaptation. Although there is evidence of performance gain when employing more advanced stain normalization like MultiStain-CycleGAN, the lack of statistical significance shows that the added signal from WSIs is subtle relative to the strong tabular predictors. Given the small effect size as well as sample size, the statistical power is small in any case, leaving room for future work. Most notably, the widely used Reinhard and Macenko stain normalization techniques induce a performance decrease, which is concordant with Voon et al.'s findings [15].

For diagnostic clinical data, the WSI signal is likely overshadowed by established clinical predictors, so the attention mechanism mostly relies on those, and the histopathological features add noise instead of signal, leading to consistently worse performance in comparison to the clinical-only baseline. Although stain normalization as well as color augmentation do show evidence of improvement over the raw extracted features, they are not able to eliminate the noise in this context.

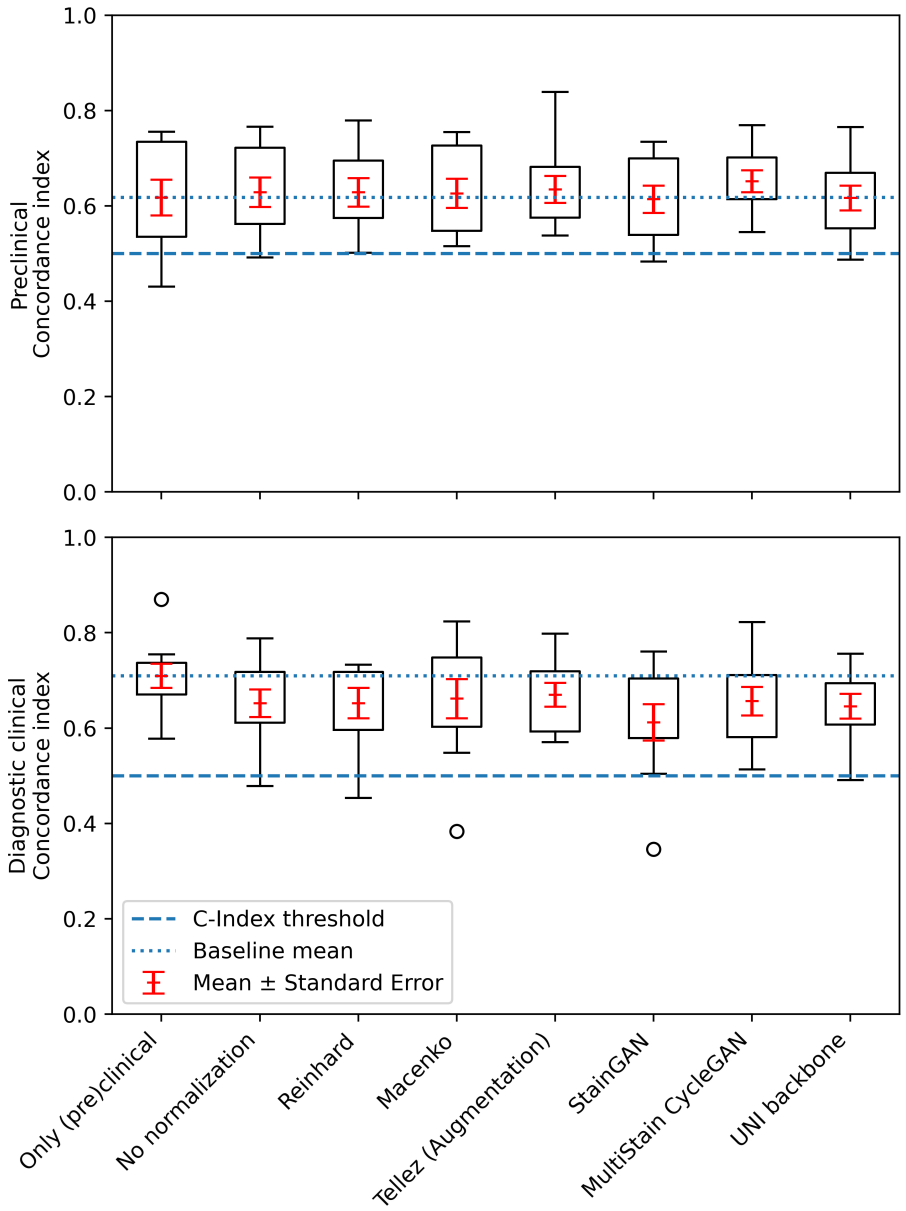

Fig. 4 Boxplot of the results. Each box represents the model performance on the held-out test set of the ten-fold cross validation. The box extends for the first quartile to the third quartile. The whiskers extend to the farthest data point within 1.5x the inter-quartile range. The C-index threshold is set at 0.5. The preclinical baseline mean is 0.618, the diagnostic clinical baseline mean is at 0.710.

Feature extractor backbone	**Clinical data**	**Stain normalization/augmentation**	**Concordance Index ± SD**	**Wilcoxon ($\mu \leq \mu_0$)**
-	Preclinical	-	0.618 ± 0.112	(baseline)
ResNet	Preclinical	-	0.629 ± 0.093	0.28
ResNet	Preclinical	Reinhard	0.629 ± 0.090	0.28
ResNet	Preclinical	Macenko	0.626 ± 0.092	0.46
ResNet	Preclinical	Tellez	0.635 ± 0.086	0.22
ResNet	Preclinical	StainGAN	0.614 ± 0.086	0.5
ResNet	Preclinical	MultiStain-CycleGAN	$\mathbf{0.652 \pm 0.069}$	0.14
UNI	Preclinical	-	0.617 ± 0.078	0.46
-	Diagnostic	-	$\mathbf{0.710 \pm 0.076}$	(baseline)
ResNet	Diagnostic	-	0.652 ± 0.087	0.97
ResNet	Diagnostic	Reinhard	0.653 ± 0.096	0.95
ResNet	Diagnostic	Macenko	0.662 ± 0.124	0.78
ResNet	Diagnostic	Tellez	0.670 ± 0.075	0.98
ResNet	Diagnostic	StainGAN	0.612 ± 0.115	0.99
ResNet	Diagnostic	MultiStain-CycleGAN	0.657 ± 0.089	0.99
UNI	Diagnostic	-	0.646 ± 0.077	0.97

Table 3 Results of model performance on the held-out test set of the ten-fold cross validation.

6 Discussion

This study investigated the impact of stain normalization techniques and the integration of clinical data on survival prognosis in breast cancer utilizing attention-based multiple instance learning from WSIs. We extended the CLAM model with early fusion of clinical features and tested several stain normalization and augmentation strategies to assess their effect on patient risk stratification.

One notable limitation of our work is the size and composition of the available dataset. Due to a limited number of censored cases, we were restricted to a maximum of ten cross-validation folds. Using more folds would have resulted in test splits with less than ten censored patients, producing unstable concordance index estimates. This highlights a persistent challenge in computational pathology: the lack of large, well-annotated, publicly available datasets encompassing high-quality WSIs as well as longitudinal data. As emphasized by Tafavvoghi et al. [47] and Song et al. [7], expanding the availability of high-quality multi-institutional data remains essential for developing robust, generalizable prognostic models and remains a challenge possibly even greater than computational constraints at this point.

Our findings indicate that the choice of stain normalization technique should be made with caution. While domain shift caused by staining variation is well recognized, our experiments show that modern CNN feature extractors, including the ResNet backbone used here, appear relatively robust to color variations. This aligns with

findings by Voon et al., suggesting that the benefit of sophisticated stain normalization may be negligible when using CNN-based pipelines trained with data augmentation. In practice, simple stain augmentation strategies may be a more resource-efficient way to account for inter-laboratory variability without introducing additional computational complexity.

The integration of clinical data through early fusion revealed an interesting pattern: adding WSIs provided modest performance gains when only preclinical data (basic anamnesis) were included, but had limited or even negative impact when stronger diagnostic variables such as TNM staging or hormone receptor status were available. This suggests that established diagnostic markers remain the dominant predictors of survival prognosis, overshadowing additional signals captured from histopathological slides. Future studies should assess more multi-modal prognostic pipelines beyond CLAM to ensure these findings hold in a broader context.

Despite significant research progress, the practical implementation of CAD and CAP systems in histopathology is limited thus far. A lack of open data, proprietary code, and missing pre-trained weights, as criticized by Wagner et al. [48], hamper independent validation and slow down both scientific progress and practical adoption. We experienced this as well when assessing recent stain normalization approaches - the vast majority did not disclose code implementations or pre-trained weights. To this end, publishing readily usable code and models should be a clear expectation for studies proposing new stain normalization or CAD and CAP pipelines.

Another challenge is bias and unintended information leakage. Site-specific staining protocols, biased patient demographics, or even scanner artifacts can be inadvertently learned by DL models, as shown by Howard et al. [37] and Taylor et al. [49]. Addressing these pitfalls requires rigorous validation on independent cohorts and transparency around training data and model behavior. Looking ahead, enhancing the interpretability of model predictions through attention-based heatmaps will strengthen trust among clinicians, facilitate practical integration into pathologists' workflows and will be a necessity for real-world clinical applications.

7 Limitations

Several limitations of this work have to be considered.

In the literature review of this work, only English studies have been included, potentially excluding others.

Regarding the data, only one dataset was considered and the zero-shot capability of the model of an entirely unseen dataset is unclear. This is due to the lack of public dataset available with follow-up annotations. Although the TCGA-BRCA dataset exhibits diverse distributions in the sense that it incorporates diagnostic slides from numerous tissue source sites with patients of various backgrounds, the ethnicity imbalance is concerning. TCGA-BRCA is also impaired by several qualitative inconsistencies that go beyond stain color variations: marker annotations, ripped or folded tissue etc. The need for high quality as well as quantity imaging for AI research has been thoroughly expressed in the literature [7]. It also remains to be seen how the

obtained results translate to other cancer types, as only invasive ductal carcinoma is considered in this study.

Concerning the model, the default settings of the CLAM model were used and no hyperparameter tuning was conducted. Since the main focus of this work was not to achieve superior overall performance but to compare stain normalization techniques among each other, this is left as future work.

Regarding the evaluation, mere risk ordering, which is assessed by the concordance index, might be impractical and of insufficient clinical interest, thus warranting more precise survival time predictions.

Overall, due to computational constraints, more normalization techniques as well as feature extractor backbones could not be investigated within the given time constraints, which remains a major point of future work.

8 Conclusion

In response to our research question, our results suggest that stain normalization techniques, as implemented here, have a limited or negligible effect on improving survival prediction performance when modern CNN-based feature extractors and stain augmentation are used.

The CNN feature extractor proved relatively robust to stain variability, with no consistent, statistically significant gains observed across different normalization strategies, at least given the small sample size. This implies that for survival prognosis tasks combining WSIs with clinical data, careful data and therefore stain augmentation during training may be sufficient to address staining variation, and complex normalization pipelines should be evaluated critically against their added implementation cost and effort. Generally, more diverse artifact-free data is preferable and should be the goal of our combined efforts.

Supplementary information. Supplementary information can be found in the subsequent appendices.

Acknowledgements. This research was supported by Prof. Fariselli and Prof. Sanavia from the Computational Biomedicine research group at the Department of Medical Sciences at the University of Turin. It was supervised by Prof. Dr. Windberger from the Heilbronn University of Applied Sciences.

Appendix A Recent applications of stain normalization in survival prognosis

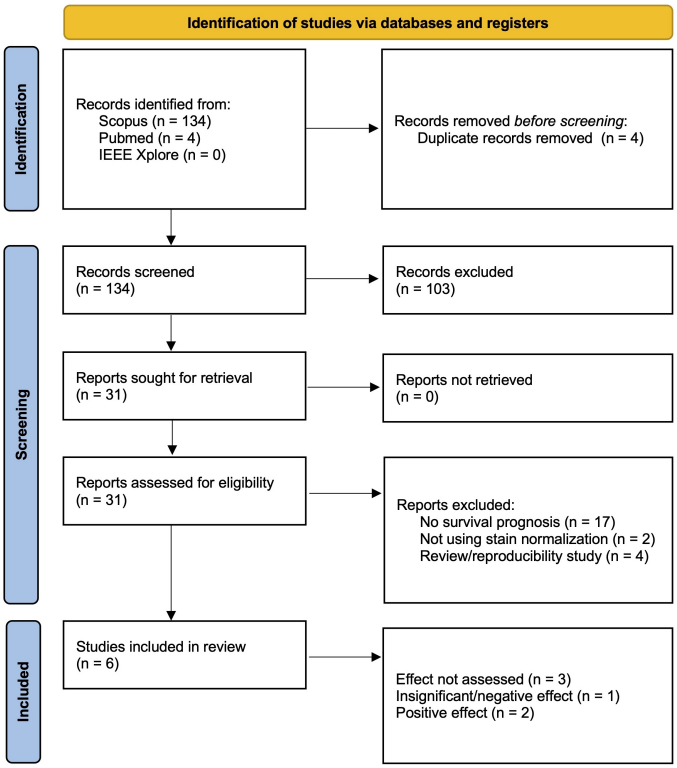

Fig. A1 PRISMA flow diagram for literature review of recent applications of stain normalization in survival prognosis using HE-stained slides.

Listing 1 Scopus search query used for literature retrieval

```
(
  TITLE-ABS-KEY (stain* AND normali*) OR
  TITLE-ABS-KEY (stain* AND augment*) OR
  TITLE-ABS-KEY (color AND normaliz*) OR
  TITLE-ABS-KEY (color AND augment*) OR
  TITLE-ABS-KEY (colour AND normalis*) OR
  TITLE-ABS-KEY (colour AND augment*)
)
AND TITLE-ABS-KEY (histopatholog*)
AND TITLE-ABS-KEY (survival OR time-to-event)
AND PUBYEAR > 1999
```

Appendix B TCGA-BRCA

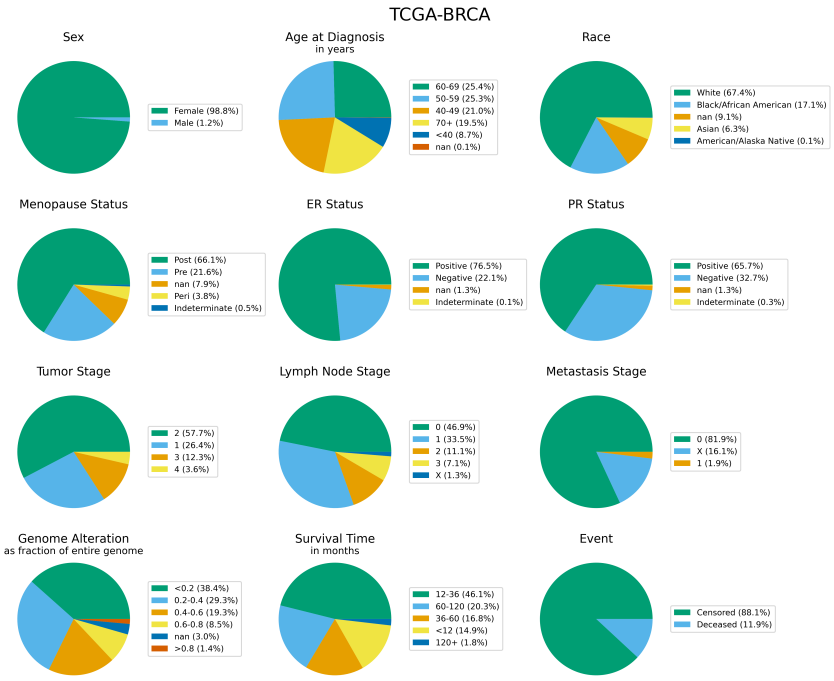

Fig. B2 Distribution of key clinical facts among included diagnostic slides of the TCGA-BRCA cohort.

References

[1] Pantanowitz, L., Sharma, A., Carter, A.B., Kurc, T., Sussman, A., Saltz, J.: Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. Journal of pathology informatics **9**(1), 40 (2018)

[2] Laak, J., Litjens, G., Ciompi, F.: Deep learning in histopathology: the path to the clinic. Nature medicine **27**(5), 775–784 (2021)

[3] Dunn, C., Brettle, D., Hodgson, C., Hughes, R., Treanor, D.: An international study of stain variability in histopathology using qualitative and quantitative analysis. Journal of Pathology Informatics, 100423 (2025)

[4] Verghese, G., Lennerz, J.K., Ruta, D., Ng, W., Thavaraj, S., Siziopikou, K.P., Naidoo, T., Rane, S., Salgado, R., Pinder, S.E., *et al.*: Computational pathology in cancer diagnosis, prognosis, and prediction–present day and prospects. The Journal of Pathology **260**(5), 551–563 (2023)

[5] Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deep survival: A deep cox proportional hazards network (2016) https://doi.org/10.48550/arXiv.1606.00931

[6] Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 544–547 (2016). IEEE

[7] Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)

[8] Alturkistani, H.A., Tashkandi, F.M., Mohammedsaleh, Z.M.: Histological stains: a literature review and case study. Global journal of health science **8**(3), 72 (2015)

[9] Hoque, M.Z., Keskinarkaus, A., Nyberg, P., Seppänen, T.: Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. Information Fusion **102**, 101997 (2024)

[10] Taqi, S.A., Sami, S.A., Sami, L.B., Zaki, S.A.: A review of artifacts in histopathology. Journal of oral and maxillofacial pathology **22**(2), 279 (2018)

[11] Kanwal, N., Pérez-Bueno, F., Schmidt, A., Engan, K., Molina, R.: The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. Ieee Access **10**, 58821–58844 (2022)

[12] Roy, S., Jain, A., Lal, S., Kini, J.: A study about color normalization methods

for histopathology images. Micron **114**, 42–61 (2018)

[13] Ke, J., Zhou, Y., Shen, Y., Guo, Y., Liu, N., Han, X., Shen, D.: Learnable color space conversion and fusion for stain normalization in pathology images. Medical Image Analysis **101**, 103424 (2025)

[14] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., Van Der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis **58**, 101544 (2019)

[15] Voon, W., Hum, Y.C., Tee, Y.K., Yap, W.-S., Nisar, H., Mokayed, H., Gupta, N., Lai, K.W.: Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images. Scientific Reports **13**(1), 20518 (2023)

[16] Boschman, J., Farahani, H., Darbandsari, A., Ahmadvand, P., Van Spankeren, A., Farnell, D., Levine, A.B., Naso, J.R., Churg, A., Jones, S.J., *et al.*: The utility of color normalization for ai-based diagnosis of hematoxylin and eosin-stained pathology images. The Journal of Pathology **256**(1), 15–24 (2022)

[17] Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2001)

[18] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: from Nano to Macro, pp. 1107–1110 (2009). IEEE

[19] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging **35**(8), 1962–1971 (2016)

[20] Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Shi, J., Xue, C.: Adaptive color deconvolution for histological wsi normalization. Computer methods and programs in biomedicine **170**, 107–120 (2019)

[21] Shaban, M.T., Baur, C., Navab, N., Albarqouni, S.: Staingan: Stain style transfer for digital histological images. In: 2019 Ieee 16th International Symposium on Biomedical Imaging (Isbi 2019), pp. 953–956 (2019). IEEE

[22] Kang, H., Luo, D., Feng, W., Zeng, S., Quan, T., Hu, J., Liu, X.: Stainnet: a fast and robust stain normalization network. Frontiers in Medicine **8**, 746307 (2021)

[23] Jain, A.K.: Fundamentals of Digital Image Processing. Prentice-Hall, Inc., ??? (1989)

[24] Khan, A.M., Rajpoot, N., Treanor, D., Magee, D.: A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE transactions on Biomedical Engineering **61**(6), 1729–1738 (2014)

[25] Zanjani, F.G., Zinger, S., Bejnordi, B.E., Laak, J.A., With, P.H.: Stain normalization of histopathology images using generative adversarial networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 573–577 (2018). IEEE

[26] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. bmj **372** (2021)

[27] Verma, J., Sandhu, A., Popli, R., Kumar, R., Khullar, V., Kansal, I., Sharma, A., Garg, K., Kashyap, N., Aurangzeb, K.: From slides to insights: Harnessing deep learning for prognostic survival prediction in human colorectal cancer histology. Open Life Sciences **18**(1), 20220777 (2023)

[28] Wang, J., Chen, J., Jiang, L., Wu, Q., Wang, D.: Predicting clear cell renal cell carcinoma survival using kurtosis of cytoplasm in the hematoxylin channel from histology slides. Journal of Oncology **2022**(1), 7693993 (2022)

[29] Yamashita, R., Long, J., Saleem, A., Rubin, D.L., Shen, J.: Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. Scientific reports **11**(1), 2047 (2021)

[30] Elforaici, M.E.A., Montagnon, E., Romero, F.P., Le, W.T., Azzi, F., Trudel, D., Nguyen, B., Turcotte, S., Tang, A., Kadoury, S.: Semi-supervised vit knowledge distillation network with style transfer normalization for colorectal liver metastases survival prediction. Medical Image Analysis **99**, 103346 (2025)

[31] Ma, B., Guo, Y., Hu, W., Yuan, F., Zhu, Z., Yu, Y., Zou, H.: Artificial intelligence-based multiclass classification of benign or malignant mucosal lesions of the stomach. Frontiers in Pharmacology **11**, 572372 (2020)

[32] Jiao, Y., Li, J., Qian, C., Fei, S.: Deep learning-based tumor microenvironment analysis in colon adenocarcinoma histopathological whole-slide images. Computer Methods and Programs in Biomedicine **204**, 106047 (2021)

[33] Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., *et al.*: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)

[34] Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult,

P., Halilovic, A., Hermsen, M., Loo, R., Vogels, R., *et al.*: 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience **7**(6), 065 (2018)

[35] Hetz, M.J., Bucher, T.-C., Brinker, T.J.: Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images. Medical Image Analysis **94**, 103149 (2024)

[36] Swinehart, D.F.: The beer-lambert law. Journal of chemical education **39**(7), 333 (1962)

[37] Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., *et al.*: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature communications **12**(1), 4423 (2021)

[38] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

[39] Yang, Z., Wei, T., Liang, Y., Yuan, X., Gao, R., Xia, Y., Zhou, J., Zhang, Y., Yu, Z.: A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. Nature Communications **16**(1), 2366 (2025)

[40] Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)

[41] Bradburn, M.J., Clark, T.G., Love, S.B., Altman, D.G.: Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. British journal of cancer **89**(3), 431–436 (2003)

[42] Faraggi, D., Simon, R.: A neural network model for survival data. Statistics in medicine **14**(1), 73–82 (1995)

[43] Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Jama **247**(18), 2543–2546 (1982)

[44] Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.-J.: On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine **30**(10), 1105–1117 (2011)

[45] Lillelund, C.M., Qi, S.-a., Greiner, R., Pedersen, C.F.: Stop Chasing the C-index: This Is How We Should Evaluate Our Survival Models (2025). https://arxiv.org/abs/2506.02075

[46] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose

foundation model for computational pathology. Nature Medicine (2024)

[47] Tafavvoghi, M., Bongo, L.A., Shvetsov, N., Busund, L.-T.R., Møllersen, K.: Publicly available datasets of breast histopathology h&e whole-slide images: a scoping review. Journal of Pathology Informatics **15**, 100363 (2024)

[48] Wagner, S.J., Matek, C., Boushehri, S.S., Boxberg, M., Lamm, L., Sadafi, A., Winter, D.J., Marr, C., Peng, T.: Built to last? reproducibility and reusability of deep learning algorithms in computational pathology. Modern Pathology **37**(1), 100350 (2024)

[49] Taylor, J., Fenner, J.: The challenge of clinical adoption—the insurmountable obstacle that will stop machine learning? BJR— Open **1**(1), 20180017 (2018)