LymphoML: An interpretable artificial intelligence-based method identifies morphologic features that correlate with lymphoma subtype

Vivek Shankar[1*] **Xiaoli Yang**[2*] **Vrishab Krishna**[1*] **Brent Tan**[3] **Oscar Silva**[3] **Rebecca Rojansky**[3] **Andrew Ng**[1] **Fabiola Valvert**[5] **Edward Briercheck**[6] **David Weinstock**[7] **Yasodha Natkunam**[3] **Sebastian Fernandez-Pol**[3†] **Pranav Rajpurkar**[4†]
[1]Department of Computer Science, Stanford University, [2]Department of Statistics, Stanford University, [3]Department of Pathology, Stanford University School of Medicine, [4]Department of Biomedical Informatics, Harvard Medical School, [5]La Liga Nacional Contra el Cáncer de Guatemala (INCAN), [6]Fred Hutchinson Cancer Research Center, [7]Dana-Farber Cancer Institute; Harvard Medical School
{vivek96,xiaoliy2,vrishab,sfernand}@stanford.edu

Abstract

The accurate classification of lymphoma subtypes using hematoxylin and eosin (H&E)-stained tissue is complicated by the wide range of morphological features these cancers can exhibit. We present LymphoML - an interpretable machine learning method that identifies morphologic features that correlate with lymphoma subtypes. Our method applies steps to process H&E-stained tissue microarray cores, segment nuclei and cells, compute features encompassing morphology, texture, and architecture, and train gradient-boosted models to make diagnostic predictions. LymphoML's interpretable models, developed on a limited volume of H&E-stained tissue, achieve non-inferior diagnostic accuracy to pathologists using whole-slide images and outperform black box deep-learning on a dataset of 670 cases from Guatemala spanning 8 lymphoma subtypes. Using SHapley Additive exPlanation (SHAP) analysis, we assess the impact of each feature on model prediction and find that nuclear shape features are most discriminative for DLBCL (F1-score: 78.7%) and classical Hodgkin lymphoma (F1-score: 74.5%). Finally, we provide the first demonstration that a model combining features from H&E-stained tissue with features from a standardized panel of 6 immunostains results in a similar diagnostic accuracy (85.3%) to a 46-stain panel (86.1%).

Keywords: model interpretability, nuclear morphology, segmentation, SHAP analysis, deep learning, digital pathology, DLBCL, Hodgkin lymphoma, Non-Hodgkin lymphoma, B-cell lymphoma, T-cell lymphoma

1. Introduction

Lymphomas are abnormal proliferations derived from lymphocytes (Jamil and Mukkamalla, 2021). The process of precisely diagnosing lymphomas requires knowledge of the clinical history (e.g. site of involvement, history of solid organ transplant) and morphologic evaluation of hematoxylin and eosin (H&E)-stained tissue by a trained pathologist (Swerdlow et al., 2016). After evaluation of the H&E-stained slide and relevant clinical information, one or a few diagnoses are deemed most likely to guide ancillary testing: immunohistochemical (IHC) stains, flow cytometry, and cytogenetic and molecular studies (Wang and Zu, 2017; Sun et al., 2016). Unlike some fields of pathology, in which a definitive diagnosis is frequently possible using H&E-stained tissue alone, for lymphoma diagnosis, IHC stains or flow cytometry are essential in most cases. This is because identifying the cell of origin for lymphomas (B-cell, T-cell, and NK cells) is essential for definitive diagnosis and treatment (Nowakowski et al., 2019), but this cannot be reliably determined based on the H&E-stained section alone. In contrast to the H&E-stained section, which is inexpensive and widely available, IHC stains and flow cytometry require costly equipment,

[*] These authors contributed equally: Vivek Shankar, Xiaoli Yang, Vrishab Krishna

[†] These authors contributed equally: Sebastian Fernandez-Pol, Pranav Rajpurkar

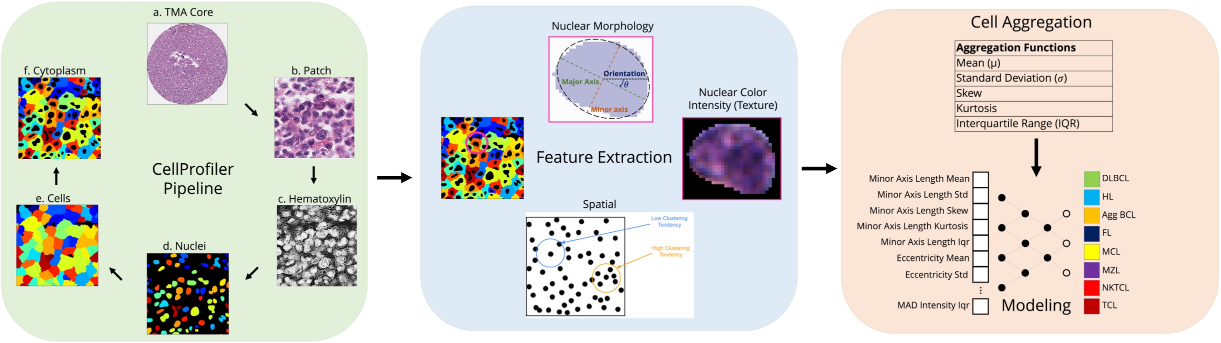


Figure 1: LymphoML Approach. We extract morphological, spatial, and textural features from segmented nuclei and cells in pathology images. We characterize the statistical distribution of each feature by computing summary metrics across the patch. These aggregated statistics are input features to train machine learning models to predict lymphoma subtypes.

expensive reagents, and trained personnel. Experienced pathologists often require fewer ancillary studies, and thus greater experience may lead to more efficient resource utilization. However, worldwide, the shortage of pathologists is so great that modest improvements in the efficiency of pathologists is unlikely to make a significant impact in reducing the costs of lymphoma diagnosis (Metter et al., 2019; Eniu et al., 2017). Thus, strategies that can help general pathologists reduce the number of ancillary studies may help to reduce the cost of lymphoma diagnosis.

Machine learning tools applied to diagnostic pathology have shown promise in analyzing H&E-stained images, achieving high accuracies ranging from 94-100% when classifying between a small number of lymphomas (ranging from 2-4 diagnostic categories: diffuse large B-cell lymphoma (DLBCL), non-DLBCL (Li et al., 2020); DLBCL, Burkitt lymphoma (BL) (Mohlman et al., 2020); DLBCL, follicular lymphoma (FL), reactive lymphoid hyperplasia (Miyoshi et al., 2020); chronic lymphocytic leukemia (CLL), FL, mantle cell lymphoma (MCL) (Janowczyk and Madabhushi, 2016; Brancati et al., 2019; Zhang et al., 2020); benign, DLBCL, BL and small lymphocytic lymphoma (Achi et al., 2019)). However, classifying between a small number of lymphoma subtypes does not reflect the full scope of complexity encompassed in pathologist workflows. Tools that can accurately distinguish among a larger number of diagnostic categories may provide greater clinical value for diagnostic pathologists in real-world settings. Furthermore, AI tools may be useful in low-middle income countries by 1) screening specimens to reduce the number of slides that require pathologist review, 2) allowing diagnoses to be made using inexpensive and widely available H&E-stained sections alone, or 3) allowing pathologists to maximize the diagnostic yield from the H&E to minimize the number of IHC-stained sections necessary to make an accurate diagnosis.

Despite prior studies achieving high performance using black-box deep-learning methods, our work highlights an interpretable, feature-engineering approach for lymphoma subtyping. We hypothesized that feature-engineering methods might out-perform deep-learning due to the limited number of labeled examples for specific lymphomas in our dataset. Additionally, our model's predictions can be explained using techniques like SHapley Additive exPlanation (SHAP) analysis (Lundberg and Lee, 2017). Model explainability is a critical component for the safety and acceptance of AI workflows in clinical practice (Evans et al., 2022).

In this work, we introduce LymphoML, an interpretable machine learning approach for lymphoma subtyping into eight diagnostic categories. LymphoML segments nuclei and cells, extracts morphological, textural, and architectural features, and aggregates them into patch-level feature vectors to train classification models. LymphoML achieves an accuracy of 64.3% on a dataset of 670 lymphoma cases from Guatemala using only H&E-stained TMA cores and demonstrates non-inferiority to hematopathologists and general pathologists. LymphoML's interpretable machine learning models outperform deep-learning: TripletNet (52.8%) and ResNet (53.5%) due to small data samples for specific diagnostic categories. Using the SHapley Additive exPlanation (SHAP) method, we find that nuclear shape features are most discriminative, especially for diffuse large B-cell lymphoma (F1-score: 78.7%), classic Hodgkin lymphoma (F1-score: 74.5%), and mantle cell lymphoma (F1-score: 71.0%). Finally, combining information from the H&E-based model with a limited set of immunohistochemical (IHC) stains results in a similar diagnostic accuracy (85.3%) as with a larger set of IHC stains (86.1%). Our work suggests a potential way to incorporate machine learning into clinical practice to reduce the number of expensive IHC stains while achieving similar diagnostic accuracy.

2. Methods

2.1. Dataset

The cases used for this study were selected and tissue microarrays (TMAs) were constructed as previously published (Valvert et al., 2021). Valvert et al. (2021) retrospectively reviewed medical records to identify formalin-fixed, paraffin-embedded (FFPE) biopsy specimens obtained at Instituto de Cancerologia y Hospital Dr. Bernardo Del Valle (INCAN) because of clinical suspicion of lymphoma between 2006 and 2018. One-half of each FFPE block was shipped to Stanford University for H&E whole-slide image (WSI) generation. Two hematopathologists reviewed the slides, selected regions of interest (ROIs), and included two cores from each sample for tissue microarray (TMA) construction. The H&E-stained TMAs were scanned at 40x magnification (0.25 µm per pixel) on an Aperio AT2 scanner (Leica Biosystems, Nussloch, Germany) in ScanScope Virtual Slide (SVS) format. Diagnoses were established based on the World Health Organization (WHO) classification (Swerdlow et al., 2016) and then binned into 8 categories: aggressive B-cell (Agg BCL), diffuse large B-cell (DLBCL), follicular (FL), classic Hodgkin (CHL), mantle cell (MCL), marginal zone (MZL), natural killer T-cell (NKTCL), or mature T-cell lymphoma (TCL). The selected categories were therapeutically driven as described in Supplemental Table 2A in Valvert et al. (2021); diagnoses that are binned together require administration of similar treatment procedures. Only a relatively small number of relevant categories were not included such as CLL, small lymphocytic leukemia, carcinoma, plasma cell neoplasm, and non-malignant cases (reactive lymphoid hyperplasia). For a full list of the categories considered in this study and the categories excluded, see Supplemental Table 2A in Valvert et al. (2021). All of the TMA blocks (seven total) were also stained for 46-different markers by IHC stains (Valvert et al., 2021). Each IHC-stained TMA was assessed by a hematopathologist to determine if the lymphoma cells were positive or negative for the marker. The complete list of cases with the associated IHC results is provided in Supplementary Table 2. The distribution of cases in each lymphoma subtype is provided in Table B.1. Of 670 FFPE biopsy specimens, 68 failed quality control (did not have sufficient tissue per core, missing ground-truth diagnoses) and were excluded from the dataset. The remaining 602 samples were split at a core-level into training, validation, and test splits with 70% of the tissue microarray (TMA) cores for training, 10% for validation to tune hyperparameters, and 20% for testing. Stratified sampling was used to proportionally represent the eight diagnostic categories in each of the training, validation, and test sets (Figure A.3).

2.2. Patch Extraction

The H&E-stained tissue cores were indicated by hematopathologists using Qupath (Bankhead et al., 2017). From each tissue core, we extracted a fixed number of non-overlapping patches at 40x magnification, starting from the top-left and proceeding until the bottom-right corner. We omitted patches that were mostly white and contained little tissue. Specifically, background was defined as pixels with saturation value less than 0.05 in HSV space, and we excluded patches where more than 95% of the pixels were background.

2.3. Nuclei and Cell Segmentation

We considered two different deep-learning based nuclear segmentation models: HoVer-Net (Graham et al., 2019) and StarDist (Schmidt et al., 2018) to segment every nucleus inside the H&E-stained TMA cores. HoVer-Net uses a neural network based on a pre-trained ResNet-50 architecture to extract image features. StarDist is powered by a pre-trained deep-learning CNN that predicts a suitable shape representation (star-convex polygon) for each cell nucleus. We normalized the input image pixel intensities to the range 0.0 to 1.0 using percentiles of 1 and 99 to clip the bottom and top 1% of pixel values to 0.0 and

1.0. Then, we ran StarDist, which operated independently on each TMA core and produced an output image segmenting all individual cell nuclei in the core. We selected StarDist as the nuclei segmentation algorithm for all our cases. We measured the agreement of HoVer-Net's and StarDist's nuclei segmentations by computing the mean Intersection over Union (mIOU) over all segmented patches. We obtained a mIOU of 0.762. Additionally, we found that the best-performing H&E-only models utilizing features extracted from StarDist achieved marginally higher top-1 accuracy (64.3%) than the best-performing models using features extracted from HoVer-Net (61.5%).

2.4. Feature Extraction

We used the per-nucleus binary segmentation masks output by StarDist to compute geometric features for each cell nucleus using methods similar to those by Vrabac et al. (2021). Using manually extracted features allowed our models to produce interpretable results and facilitated identification of features that were most important in driving the classification using SHAP (SHapley Additive exPlanations, described below) (Lundberg and Lee, 2017). We calculated features such as Feret diameters, convex hull area of the segmented nucleus, and derived geometric features including measures of circularity, elongation, and convexity. To obtain a richer feature set, we used CellProfiler (Carpenter et al., 2006), an open-source tool for analyzing biological images, to extract quantitative features of the morphology, color intensity, and texture of segmented nuclei and cells. We constructed an image analysis pipeline in CellProfiler consisting of modules to process the H&E cores, identify nuclear and cell boundaries, and measure features of the identified objects (Figure 1). First, a color deconvolution was performed on each patch to create separate hematoxylin and eosin-stained images in grayscale. Next, we ran StarDist on the hematoxylin image to produce a binary mask segmenting the nuclei. The nuclei were subsequently used as a reference to identify secondary objects such as the cells and cytoplasm. Finally, we extracted size, shape (e.g. bounding box area, minor axis length), color intensity (e.g. mean intensity, integrated intensity), and textural features from the detected cells and nuclei. The full list of features extracted by CellProfiler is provided in Table B.3 (Stirling et al., 2021). To obtain a single feature vector for each patient, each of the features was aggregated across all nuclei in a patch by their mean, standard deviation, skew, kurtosis, and percentiles, yielding a total of 1595 features for each patient.

Spatial Relationship Features. To model spatial relationships between nuclei, we considered architectural features from two sources: 1) CPArch: features that contain architectural information provided by CellProfiler (BoundingBoxMaximum, Center), and 2) CT: spatial features representing clustering tendency (CT) using Ripley's K function. We followed the steps described in Subramanian et al. (2018) to compute CT. We used centroid coordinates (in pixels) to define cell locations in each patch and computed values of the self-K function at different radii. The optimal radii range was determined by cross-validation. The resulting vector consisting of self-K function values at each radius was used as the patch's CT feature. When performing lymphoma subtype predictions, we concatenated CPArch and CT vectors directly with the rest of the features.

2.5. Models

We used LightGBM (Ke et al., 2017), a tree-based machine learning algorithm that employs a gradient boosting framework. We handled class-imbalance by preserving the label distribution when splitting the dataset and made sure patches from the same patient were in the same data split. To correct the bias induced by class-imbalance during model training, we used focal loss (Lin et al., 2017) and turned on 'balanced' mode in LightGBM to adjust weights inversely proportional to class frequencies in the input data. We used 5-fold cross-validation for all experiments. We experimented with hyperparameter tuning on the number of leaves, maximum depth, and number of epochs for gradient-boosting models.

For deep-learning models, we divided cores into patches of 224x224 pixels with 50% overlap data augmentation and filtered patches as described in "Patch Extraction." Patch pixels were normalized to have mean 0, variance 1. We fine-tuned two open-source models pretrained on H&E patches − a ResNet-50 self-supervised on several tasks and cancers with H&E and IHC-stained slides (He et al., 2016) and a specialized TripletNet architecture pretrained on CAMELYON16 (dataset of breast cancer H&E WSIs) (Srinidhi et al., 2022). We experimented with hyperparameter tuning on the learning rate (in the range: 1e-2, 1e-3, 1e-4, 1e-5) and unfreezing different numbers of layers of the pre-trained Triplet-

Net and ResNet while fine-tuning. The deep-learning results we reported use the best identified hyperparameters (learning rate: 0.001, allowing weights in all layers to update during fine-tuning) on the validation set. Focal loss was used to handle class-imbalance with normalized weights generated from label proportions on the training set and a gamma parameter of 2.0. The model was updated by an Adam Optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 and batch size of 128 for 100 epochs.

2.6. Feature Importances

We used the SHapley Additive exPlanation (SHAP) method to quantify the impact of each feature on the trained model (Lundberg and Lee, 2017). The SHAP method explains prediction by allocating credit among input features; feature credit is calculated using Shapley Values as the change in the expected value of the model's predicted score for a label when a feature is present versus absent. We also grouped related morphological features into different categories and ran SHAP on the feature groups. We summed the raw SHAP values within each group to estimate the group's importance.

2.7. Evaluation

We assessed the performance of our models (index test) in predicting the ground-truth WHO diagnosis (reference standard) for each case. The assessors of the reference standard reviewed H&E slides and IHC results to classify each specimen according to the WHO classification (Valvert et al., 2021). We additionally compared our models to human-benchmark pathologists. The H&E-stained TMA cores/WSIs were reviewed by hematopathologists who were blinded to all other clinical data including immunohistochemical stains and the final histopathologic diagnosis (reference standard result). Other than the H&E-stained tissue, no additional clinical information were provided to the human-benchmark pathologist or to our models.

We analyzed model performance by top-1 accuracy, F1 score, AUROC, sensitivity, and specificity. Top-1 accuracy, the proportion of examples for which the predicted label matches the target label, provides a direct measure of model/pathologist performance and was used in several prior works including Li et al. (2020) and Steinbuss et al. (2021). We calculated metrics for each label individually and their weighted average across all labels. We weight by the support, the number of true instances for each label, to account for class imbalance. Let k represent the total number of labels, n represent the total number of examples, and $|y_i|$ represent the number of examples with label i. We compute the weighted F1 score:

$$\text{Weighted\_F1\_Score} = \frac{1}{n}\sum_{i=1}^{k} |y_i| * \text{F1\_Score}_i \qquad (1)$$

We followed a similar procedure to calculate weighted sensitivity, specificity, and AUROC metrics. We computed 95% CIs using non-parametric bootstrapping from 1,000 bootstrap samples. These metrics are summarized in Table B.6 for models using features extracted from H&E-stained tissue only. We performed paired t-tests checking for any significant differences and equivalence (TOST) relationships by comparing the top-1 accuracy of the Best H&E Model with pathologists on H&E-stained TMA cores and WSIs. The two-tailed paired t-test between the Best H&E Model and Pathologist checks whether the mean difference for a particular metric (e.g. top-1 accuracy) between the two methods is zero. The significance level is 5%. Let $\mu_{\text{Best H\&E Model}}$ represent the mean of a metric (e.g. mean top-1 accuracy) for the Best H&E Model and $\mu_{\text{Pathologist}}$ represent the mean of a metric for the Pathologist. TOST consists of two one-tailed paired t-tests with the following null hypothesis:

$$|\mu_{\text{Best H\&E Model}} - \mu_{\text{Pathologist}}| \geq +\Delta \qquad (2)$$

Statistical test results are summarized in Table B.10. This study was conducted in compliance with STARD guidelines (Supplementary Table 1).

3. Results

3.1. Nuclear Morphology

We first tested whether nuclear shape features had higher diagnostic yield than nuclear texture or cytoplasmic features for classifying lymphoma subtypes. The model using only nuclear features achieved 59.7% ([51.2%, 68.2%]) top-1 test accuracy, while models using nuclear texture or cytoplasmic features alone achieved slightly lower accuracy (Table B.6). Adding nuclear texture or cytoplasmic features to the nuclear shape features marginally improved performance by 1-2%. We analyzed model performance by lymphoma subtype using per-class F1 scores. Nuclear features were most discriminative for diffuse large B-cell lymphoma (DLBCL) (F1 score: 76.2%), classic Hodgkin lymphoma (CHL) (F1 score: 65.3%), and mantle cell lymphoma (MCL) (F1 score: 51.6%) (Table B.7). To extract morphologic features from H&E-stained images, we used the data processing, feature extraction, and modeling workflow described in Figure A.1.

Feature Importances. We utilized the SHAP method to investigate whether area shape features played the biggest role in classifying lymphoma subtypes among all nuclear features. We reported the resulting Shapley values for the top 20 nuclear features on individual predictions. We also trained a parsimonious model using only the top eight most impactful features for each class as determined by SHAP; the resulting model achieved a top-1 test accuracy of 61.2% ([53.4%, 69.0%]) using just 10% of all features. The majority of the top 20 nuclear features were area shape features such as mean radius, minor axis length, maximum feret diameter, solidity, orientation, maximum radius length, and nuclei area.

For select features, we analyzed their ability to distinguish patients with specific lymphoma subtypes. For example, the MinorAxisLength parameter was significantly different between cases of DLBCL and MCL (Figure A.2). The minor axis length (MAL) is the length (in pixels) of the ellipse that has the same normalized second central moments as the nucleus region (Stirling et al., 2021). DLBCL generally has cells with a larger minor axis length, consistent with the WHO definition of DLBCL as a B-cell lymphoma with large cells (Swerdlow et al., 2016). We grouped related morphological features into different categories and ran SHAP on the resulting feature groups. The nuclear size feature group had the largest mean absolute SHAP value (Figure A.4), suggesting that of all nuclear features, size features were most helpful for classifying DLBCL, CHL, and MCL.

3.2. Nuclear Architecture

We hypothesized that incorporating architectural features such as clustering tendency among nuclei would improve accuracy on certain types of lymphoma over nuclear morphological features alone. The best model utilizing architectural and nuclear features was “Nuclear Morphological + Cytoplasm + Intensity + CPArch” (referred to as “Best H&E Model” below). It achieved 64.3% ([55.7%, 72.9%]) top-1 accuracy, the highest accuracy among all models using H&E features, though statistical tests did not suggest it being significantly different from the Nuclear Morphological Model. Although Best H&E Model, which uses all inputs except CT, slightly outperforms the model using all available features, the difference between their performances is not statistically significant (Figure 2). We examined Best H&E Model's per-class performance to find that it achieved 71.0% ([55.0%, 87.0%]) F1 score in predicting MCL, 19.4% higher than the Nuclear Morphological Model's F1 score for MCL (51.6% [32.2%, 71.0%]), though this difference did not achieve statistical significance.

3.3. Comparison to Black-Box Features

We hypothesized that the interpretable machine learning models produced by LymphoML would outperform deep-learning methods (TripletNet and ResNet) used in prior studies given the scarcity of labeled examples. There was no significant difference between the test performance of the two deep-learning methods. ResNet, which achieved better test accuracy than TripletNet, was significantly inferior to the Best H&E Model. It was also worse than the Nuclear Morphological Model in both test accuracy and F1 score by a margin of ~5% (top-1 test accuracy of TripletNet: 52.8% [44.2%, 61.4%], top-1 test accuracy of ResNet: 53.5% [44.8%, 62.2%]). For each class, the baseline generally had better performances.

3.4. Pathologist Comparison

We compared the performance of the Best H&E Model from Section 3.2 with pathologists on H&E-stained TMA cores and WSIs (Figure 3). Best H&E Model's test accuracy (64.3% [55.7%, 72.9%]) surpassed all pathologists (Table 1). Statistical tests verified that Best H&E Model was non-inferior to the General Pathologist on WSIs and Hematopathologist 1 on TMAs, and there was no evidence of its performance being statistically different from any pathologist's performance (Table B.10). A similar conclusion can be derived from other metrics (sensitivity, specificity, and AUROC). We compared the per-class performance of our methods. On one hand, compared to pathologists' predictions, the Best H&E Model failed to effectively identify any case in MZL and TCL while pathologists generally achieved 30% and 23.5% F1 scores in diagnosing MZL and TCL respectively. On the other hand, Best H&E Model achieved a 71.0% F1 score ([55.0%, 87.0%]) in MCL, surpassing all hematopathologists by a margin >18% (Table 2) and statistically superior to Hematopathologist 1 on H&E TMAs and Hematopathologist 3 on WSIs. Best H&E Model achieves consistent and better performance than hematopathologists only for the 3 categories of DLBCL, CHL, and MCL for which we have a sufficiently large number of cases in our cohort.

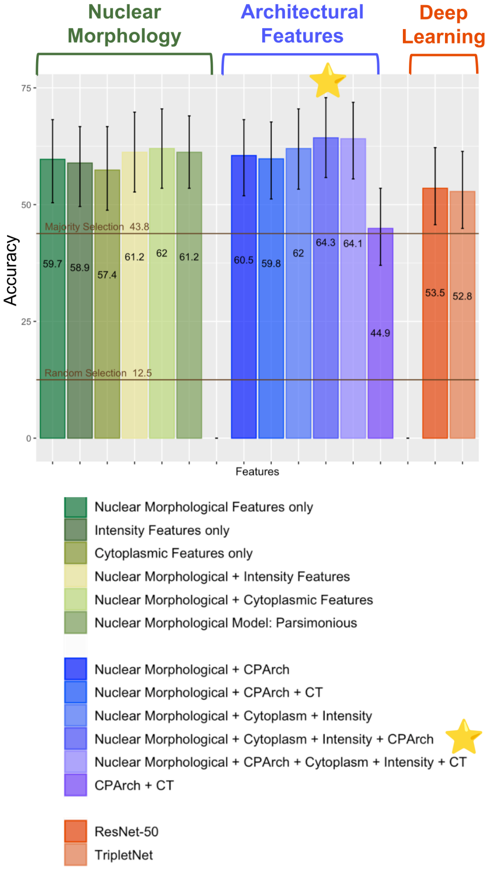


Figure 2: Performance summary for all H&E models with Majority Selection line representing a baseline that always predicts the most prevalent class and Random Selection line representing a baseline that performs random classification. The best performing model ("Best H&E Model") is marked with a star.

3.5. Additional Stains

The ground-truth diagnosis for our cohort was based on review by hematopathologists of the H&E-stained whole-slide images and a panel of 46 immunohistochemical (IHC) stains that were performed on all cases. For each case and immunostain, a hematopathologist scored the lymphoma cells as positive or negative for the immunostain (e.g. CD30-positive) or "cannot interpret." We set out to determine the diagnostic accuracy of using this IHC information alone, using a limited set of IHC stain results, and using a limited set of IHC stain results in conjunction with the H&E-based model.

We considered six candidate immunostains based on our impression of markers that would provide the highest yield considering the diagnostic categories in our cohort: CD10, CD20, CD3, EBV-ISH, BCL1/cyclin D1, and CD30. For each immunostain, the pathologist's score was included as an additional categorical feature to the Best H&E Model. We also grouped lymphoma subtypes into five categories that are similar in terms of clinical behavior and therapeutic approaches: B-cell lymphomas (DLBCL, Agg BCL), CHL, FL and MZL, MCL, and T-cell lymphomas (NKTCL, TCL). We evaluated the accuracy and F1 scores of models using features extracted from H&E stains along with different combinations of immunostain indicator features. The baseline model using 46 immunostains (and no H&E) achieved a top-1 accuracy of 86.1% ([80.0%, 92.2%]). Using the six selected immunostains alone, without the H&E, the model achieved an accuracy of 75.2% ([68.2%, 82.2%]), statistically inferior to the model using all 46 immunostains (Figure 3). The Best H&E Model augmented by six immunostains (CD10, CD20, CD3, EBV-ISH, BCL1, CD30) achieved an accuracy of 85.3% ([79.9%, 90.7%]) and showed no evidence of difference from the model using all 46 immunostains.

4. Discussion

In this study, we assessed the performance of interpretable and deep-learning approaches in the classification of eight lymphoma categories using a cohort of 670 lymphoma cases. Using only H&E-stained TMA material, LymphoML achieves performance comparable to that of experienced hematopathologists reviewing only the H&E-stained tissue. The highest performance was achieved using an interpretable model. Combining information from the H&E-based model with a limited set of IHC stains resulted in a similar diagnostic accuracy as with a much larger set of IHC stains. It is notable that our interpretable models, developed on a limited volume of tissue, achieved the same diagnostic accuracy as hematopathologists using whole-slide images. Our study suggests that computational tools can extract more diagnostic information from less tissue than a pathologist. The growing

Method	Accuracy	Sensitivity	Specificity	AUROC	F1 Score
Hematopathologist 1 on H&E TMAs	56.1 ± 8.1	56.5 ± 9.7	82.1 ± 5.4	N/A	53.5 ± 8.5
Hematopathologist 2 on H&E TMAs	60.1 ± 7.5	60.5 ± 8.2	82.8 ± 5.4	N/A	58.8 ± 8.4
Hematopathologist 3 on WSIs	63.5 ± 7.4	63.9 ± 9.9	92.9 ± 2.8	N/A	66.0 ± 8.2
General Pathologist on WSIs	56.1 ± 7.4	55.8 ± 9.7	93.2 ± 1.9	N/A	65.1 ± 7.3
Best H&E Model	64.3 ± 8.6	66.9 ± 6.0	88.7 ± 2.6	85.9 ± 2.9	58.5 ± 9.5

Table 1: Overall class-weighted performance metrics of Best H&E Model vs pathologists using TMAs/WSIs.

Method	DLBCL	HL	Agg BCL	FL
Hematopathologist 1 on H&E TMAs	73.3 ± 7.8	63.8 ± 14.9	25.0 ± 25.0	35.7 ± 23.6
Hematopathologist 2 on H&E TMAs	73.3 ± 7.4	86.4 ± 9.4	0.0	42.9 ± 21.6
Hematopathologist 3 on WSIs	82.1 ± 6.6	86.4 ± 9.4	0.0	58.3 ± 24.5
General Pathologist on WSIs	67.8 ± 9.3	80.0 ± 12.3	0.0	54.5 ± 21.4
Best H&E Model	78.7 ± 7.7	74.5 ± 13.4	0.0	31.6 ± 24.8
Method	**MCL**	**MZL**	**NKTCL**	**TCL**
Hematopathologist 1 on H&E TMAs	36.4 ± 23.6	28.6 ± 28.6	0.0	23.5 ± 23.5
Hematopathologist 2 on H&E TMAs	43.5 ± 23.2	17.4 ± 17.4	26.7 ± 26.7	0.0
Hematopathologist 3 on WSIs	20.0 ± 20.0	30.8 ± 30.7	70.0 ± 20.0	23.5 ± 23.5
General Pathologist on WSIs	52.2 ± 23.7	50.0 ± 38.9	1.0	23.5 ± 23.5
Best H&E Model	71.0 ± 16.0	0.0	25.0 ± 25.0	0.0

Table 2: Per-class F1 score comparison of Best H&E Model to pathologists using TMAs/WSIs.

use of needle core biopsies in clinical practice makes judicious use of tissue even more critical. Computational tools that maximize the diagnostic yield of the H&E-stained slide could potentially reduce the number of ancillary stains and thus avoid the need for repeat biopsy or an excisional biopsy.

Feature engineering effective on limited data. Prior studies that applied machine learning methods to lymphoma diagnosis have achieved accuracies of 94% to 100%, but it is not clear how these tools can be employed in the real-world given the limited number of diagnostic categories studied. Our study was performed on a cohort that contains 2- to 4-fold more diagnostic categories than prior studies. Though our cohort contains the second-largest number of cases thus far used to develop machine learning tools for lymphoma diagnosis, the number of cases is within the same order of magnitude as that used in other studies. In real-world settings with a large number of possible diagnoses, deep-learning methods will require a proportionate increase in the number of cases to perform well. With a limited number of examples for specific diagnostic categories, models can only learn a small range of the wide morphological variability present in these subtypes. Therefore, feature-engineering approaches may yet provide superior diagnostic yield when the number of cases per diagnosis is insufficient to use deep-learning. Furthermore, we note that deep-learning models will have comparable performance to prior work (e.g. DLBCL vs non-DLBCL by Li et al. (2020)) if they are limited to the same number of classes and provided a sufficiently large number of samples per class.

Nuclear features align with pathologist review. Pathologists use features similar to the nuclear shape features identified by SHAP including nuclear-to-cytoplasmic ratio, nuclear contour irregularities, and nuclear size as one of many clues to determine if cells are normal or malignant. DLBCL, by definition, consists of sheets of large B-cells, with 'large' defined as nuclei that are at least the size of a histiocyte nucleus or two-times the size of a normal lymphocyte. Consistent with this definition used to render the ground-truth diagnosis, our interpretable model showed that nuclear size features differed between DLBCL and other lymphoma subtypes. Specific features that provided the greatest diagnostic yield included mean radius, minor axis length, maximum Feret diameter, solidity, and orientation.

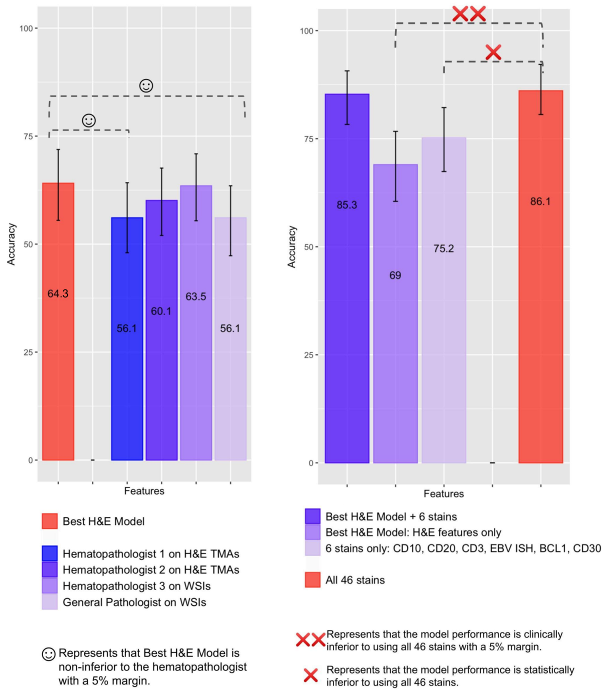


Figure 3: Performance comparison between hematopathologists and the best-performing H&E model (left). Performance comparisons of models using features extracted from H&E and selected immunostains (right).

Combining assessment with immunostains can be cost-effective. In our study, we provide the first demonstration that a model that incorporates immunophenotypic features from six selected IHC stains and features extracted from H&E-stained sections achieves the same diagnostic accuracy as a model that uses a larger number of immunostains. As H&E-stained slides are cheaper than IHC stains by at least an order of magnitude, extracting the maximal diagnostic yield from H&E can reduce the number of IHC stains ordered and costs without a reduction in diagnostic accuracy. Additionally, we note that pathologists usually order a custom panel of IHC stains for each case. In our study, we showed that models combining H&E features with a standard set of six IHC stains can arrive at the correct diagnosis (85.3%). Standardizing IHCs reduces variability in practice and improves cost effectiveness.

Strengths. The best LymphoML models achieved top-1 test accuracy, sensitivity, and specificity equivalent to that of experienced hematopathologists. Using feature importance analysis, our computational methods help pathologists better understand the primary characteristics of the lesion that contribute to the model's prediction. Our study cohort from Guatemala, a population that is not represented in current digital pathology datasets, will prove valuable in the effort to build diverse datasets for computational tools in medicine. Most prior works focused on the use of whole-slide images, often with expensive, manually-generated patch-level annotations. WSIs often require manual annotations because these larger images usually contain both cancerous and normal surrounding tissue. Here, we use TMAs, which do not require expensive manual annotations because cores are already enriched for lymphoma. Using TMAs, computational tools are more cost-effective and usable in low-middle income countries where acquisition of labeled data can prove costly.

Limitations. First, the data were collected and processed in a single institution by a single slide scanner. We should evaluate the model's generalizability on cohorts from other institutions collected using different technical setups and slides scanned on different machines. Next, TMAs capture only a small portion of the full tumor volume and are much smaller than WSIs (Beck et al., 2011). Thus, we could have likely trained more powerful models by analyzing WSIs. The heavy class imbalance meant that only a small number of examples were available for Agg BCL, MZL, TCL (<10 patients), which can only display a small subset of the wide morphological variability present in these subtypes. Finally, to implement LymphoML in a clinical setting, pathologists would have to suspect one of the validated diagnostic categories, select a region of interest, and then the model would render a favored diagnosis. Future studies with more diagnostic categories, and a more diverse set of background tissues may someday enable automated identification of target lesion(s) and subsequent diagnostic categorization.

Data and Code Availability

The datasets used and/or analyzed during the current study are available from Sebastian Fernandez-Pol (sfernand@stanford.edu) on reasonable request. We are acquiring a Data Use Agreement (DUA) to be able to share the raw data. Our code is made available on Github at this link.

Author Contributions

V.S., X.Y., V.K., S.F., and P.R. developed the concept and design; V.S., X.Y., V.K., S.F., P.R., B.T., O.S., R.R., F.V., E.B., D.W., and Y.N. performed acquisition, analysis, or interpretation of data; A.N., S.F., and P.R. provided supervision. V.S., X.Y., V.K., S.F., and P.R. drafted the manuscript, and all authors provided critical revision of manuscript for important intellectual content.

Ethics

This study was approved by the institutional review boards of the Dana-Farber Cancer Institute and Stanford University and the ethics committee of La Liga Nacional Contra el Cáncer.

References

Hanadi El Achi, Tatiana Belousova, Lei Chen, Amer Wahed, Iris Wang, Zhihong Hu, Zeyad Kanaan, Adan Rios, and Andy N D Nguyen. Automated diagnosis of lymphoma with digital pathology images using deep learning. *Ann. Clin. Lab. Sci.*, 49 (2):153–160, March 2019.

Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, Jacqueline A James, Manuel Salto-Tellez, and Peter W Hamilton. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.*, 7 (1):16878, December 2017.

Andrew H Beck, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J van de Vijver, Robert B West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.*, 3(108): 108ra113, November 2011.

Nadia Brancati, Giuseppe De Pietro, Maria Frucci, and Daniel Riccio. A deep learning approach for breast invasive ductal carcinoma detection and lymphoma Multi-Classification in histological images. *IEEE Access*, 7:44709–44720, 2019.

Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100, October 2006.

Alexandru E Eniu, Yehoda M Martei, Edward L Trimble, and Lawrence N Shulman. Cancer care and control as a human right: Recognizing global oncology as an academic field. *Am Soc Clin Oncol Educ Book*, 37:409–415, 2017.

Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. The explainability paradox: Challenges for xAI in digital pathology. *Future Gener. Comput. Syst.*, 133:281–296, August 2022.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.*, 58:101563, December 2019.

Shilpa Gupta, Ruchika Gupta, Sompal Singh, Kusum Gupta, and Madhur Kudesia. Nuclear morphometry and texture analysis of b-cell non-hodgkin lymphoma: utility in subclassification on cytosmears. *Diagn. Cytopathol.*, 38(2):94–103, February 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Ayesha Jamil and Shiva Kumar R Mukkamalla. Lymphoma. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), September 2021.

Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.*, 7:29, July 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

C Lesty, M Raphael, L Nonnenmacher, V Leblond-Missenard, A Delcourt, A Homond, and J L Binet. An application of mathematical morphology to analysis of the size and shape of nuclei in tissue sections of non-hodgkin's lymphoma. *Cytometry*, 7 (2):117–131, March 1986.

Dongguang Li, Jacob R Bledsoe, Yu Zeng, Wei Liu, Yiguo Hu, Ke Bi, Aibin Liang, and Shaoguang Li. A deep learning diagnostic platform for diffuse large b-cell lymphoma with high accuracy across multiple hospitals. *Nat. Commun.*, 11(1):6004, November 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

David M Metter, Terence J Colgan, Stanley T Leung, Charles F Timmons, and Jason Y Park. Trends in the US and canadian pathologist workforces from 2007 to 2017. *JAMA Netw Open*, 2(5):e194337, May 2019.

Hiroaki Miyoshi, Kensaku Sato, Yoshinori Kabeya, Sho Yonezawa, Hiroki Nakano, Yusuke Takeuchi, Issei Ozawa, Shoichi Higo, Eriko Yanagida, Kyohei Yamada, Kei Kohno, Takuya Furuta, Hiroko Muta, Mai Takeuchi, Yuya Sasaki, Takuro Yoshimura, Kotaro Matsuda, Reiji Muto, Mayuko Moritsubo, Kanako Inoue, Takaharu Suzuki, Hiroaki Sekinaga, and Koichi Ohshima. Deep learning shows the capability of high-level computer-aided diagnosis in malignant lymphoma. *Lab. Invest.*, 100(10):1300–1310, October 2020.

Jeffrey S Mohlman, Samuel D Leventhal, Taft Hansen, Jessica Kohan, Valerio Pascucci, and Mohamed E Salama. Improving augmented human intelligence to distinguish burkitt lymphoma from diffuse large B-Cell lymphoma cases. *Am. J. Clin. Pathol.*, 153(6):743–759, May 2020.

Grzegorz S Nowakowski, Tatyana Feldman, Lisa M Rimsza, Jason R Westin, Thomas E Witzig, and Pier Luigi Zinzani. Integrating precision medicine through evaluation of cell of origin in treatment planning for diffuse large b-cell lymphoma. *Blood Cancer J.*, 9(6):48, May 2019.

Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 265–273. Springer International Publishing, 2018. doi: 10.1007/978-3-030-00934-2_30. URL https://doi.org/10.1007%2F978-3-030-00934-2_30.

Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med. Image Anal.*, 75:102256, January 2022.

Georg Steinbuss, Mark Kriegsmann, Christiane Zgorzelski, Alexander Brobeil, Benjamin Goeppert, Sascha Dietrich, Gunhild Mechtersheimer, and Katharina Kriegsmann. Deep learning for the classification of non-hodgkin lymphoma on histopathological images. *Cancers*, 13:2419, 05 2021. doi: 10.3390/cancers13102419.

David R Stirling, Madison J Swain-Bowden, Alice M Lucas, Anne E Carpenter, Beth A Cimini, and Allen Goodman. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22(1):433, September 2021.

Vaishnavi Subramanian, Weizhao Tang, Benjamin Chidester, Jian Ma, and Minh N Do. Integration of spatial distribution in Imaging-Genetics. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 245–253. Springer International Publishing, 2018.

Ruifang Sun, L Jeffrey Medeiros, and Ken H Young. Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era of precision medicine. *Mod. Pathol.*, 29(10):1118–1142, October 2016.

Steven H Swerdlow, Elias Campo, Stefano A Pileri, Nancy Lee Harris, Harald Stein, Reiner Siebert, Ranjana Advani, Michele Ghielmini, Gilles A Salles, Andrew D Zelenetz, and Elaine S Jaffe. The 2016 revision of the world health organization classification of lymphoid neoplasms. *Blood*, 127(20): 2375–2390, May 2016.

Fabiola Valvert, Oscar Silva, Elizabeth Solórzano-Ortiz, Maneka Puligandla, Marcos Mauricio

Siliézar Tala, Timothy Guyon, Samuel L Dixon, Nelly López, Francisco López, César Camilo Carías Alvarado, Robert Terbrueggen, Kristen E Stevenson, Yasodha Natkunam, David M Weinstock, and Edward L Briercheck. Low-cost transcriptional diagnostic to accurately categorize lymphomas in low- and middle-income countries. *Blood Adv*, 5(10):2447–2455, May 2021.

Damir Vrabac, Akshay Smit, Rebecca Rojansky, Yasodha Natkunam, Ranjana H Advani, Andrew Y Ng, Sebastian Fernandez-Pol, and Pranav Rajpurkar. DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set. *Sci Data*, 8(1):135, May 2021.

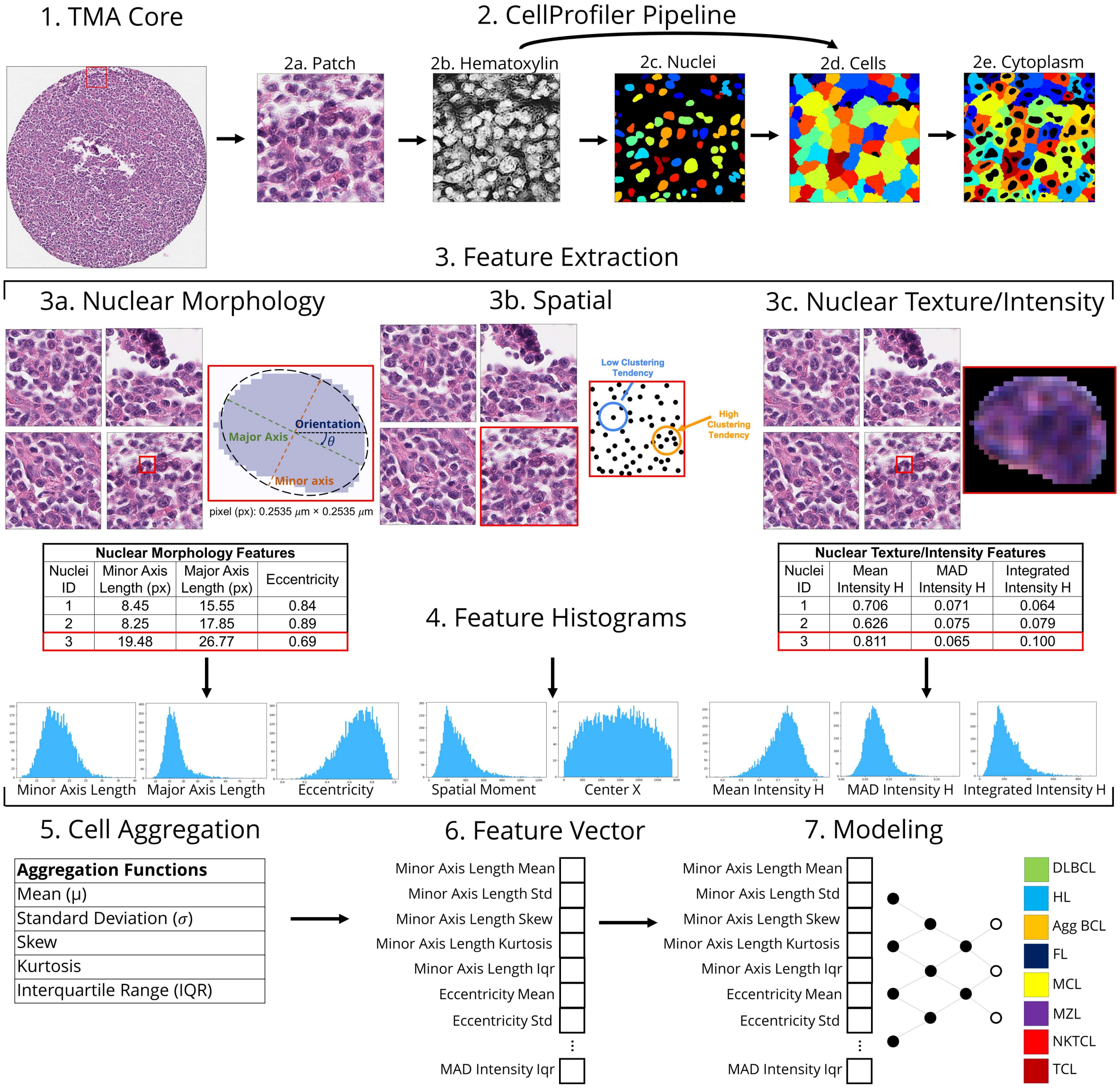
Huan-You Wang and Youli Zu. Diagnostic algorithm of common mature B-Cell lymphomas by immunohistochemistry. *Arch. Pathol. Lab. Med.*, 141(9): 1236–1246, September 2017.

Wei-Hsiang Yu, Chih-Hao Li, Renching Wang, Chao-Yuan Yeh, and Shih-Sung Chuang. Machine learning based on morphological features enables classification of primary intestinal t-cell lymphomas. *Cancers*, 13:5463, 10 2021. doi: 10.3390/cancers13215463.

Jianfei Zhang, Wensheng Cui, Xiaoyan Guo, Bo Wang, and Zhen Wang. Classification of digital pathological images of non-hodgkin's lymphoma subtypes based on the fusion of transfer learning and principal component analysis. *Med. Phys.*, 47 (9):4241–4253, September 2020.

Menglei Zhu, Michael Hazoglou, Anyi Li, and Ahmet Dogan. Automatic triaging of hematopathology tissue specimens by neural network on whole slide image (wsi). *Laboratory Investigation*, 102: 1058–59, March 2022.

Appendix A. Supplementary Figures



Nuclear Morphology Features			
Nuclei ID	Minor Axis Length (px)	Major Axis Length (px)	Eccentricity
1	8.45	15.55	0.84
2	8.25	17.85	0.89
3	19.48	26.77	0.69

Nuclear Texture/Intensity Features			
Nuclei ID	Mean Intensity H	MAD Intensity H	Integrated Intensity H
1	0.706	0.071	0.064
2	0.626	0.075	0.079
3	0.811	0.065	0.100

Aggregation Functions
Mean (μ)
Standard Deviation (σ)
Skew
Kurtosis
Interquartile Range (IQR)

Figure A.1: LymphoML Approach. Patches of a fixed size are extracted from each TMA core. We used the StarDist algorithm to produce a nuclei segmentation mask. Using the H image and the nuclei segmentation mask together, we used CellProfiler to identify the cell boundaries. We identified the cytoplasm by "subtracting" the nuclei objects from the cell objects. For each identified nucleus, cell, and cytoplasm, the following groups of features were extracted and measured: (a) nuclear morphology features, (b) spatial/architectural features (c) texture/intensity features. For each measurement obtained, the mean, standard deviation, skew, kurtosis, and interquartile range (IQR) were calculated for the entire population of objects present in a patch. The aggregated features were packed into a patch-level feature vector. Using the feature vector as input, models were trained to predict the most likely lymphoma subtype for the patch. The plurality vote across the patch-level model predictions was used to obtain the final core-level prediction.

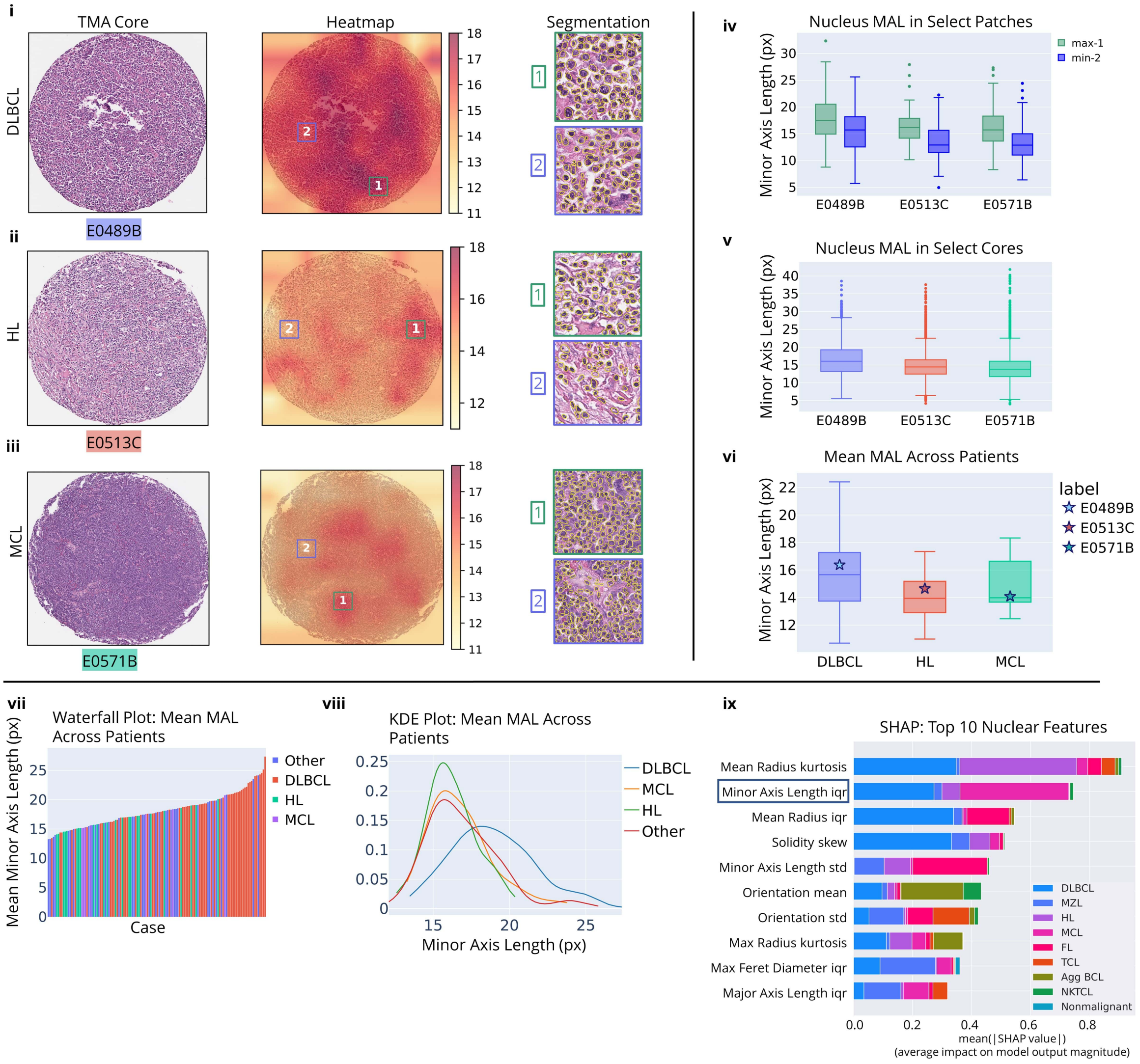


Figure A.2: Comparison of Minor Axis Length (MAL) across lymphoma subtypes. For selected cases from (i) DLBCL, (ii) CHL/HL, and (iii) MCL, we plot a heatmap superimposed on the tissue microarray (TMA) cores showing the variability in nuclei minor axis length across the core. We also mark the patches with the largest (1) and smallest (2) mean MAL. For each patch, we display the segmented nuclei using StarDist. In (iv), we show the distribution of nuclei MAL in the maximum (1) and minimum (2) patches for each case. In (v), we show the distribution of nuclei MAL in the overall cores. In (vi), we show the distribution of mean MAL across cores in the entire TMA. DLBCL cases generally have the largest nuclei and greater variability in nuclei size than either CHL or MCL cases. In (vii), we display a waterfall plot of the mean minor axis length across patients. For each case/patient, we plot the mean of the top five patches with the highest minor axis length. DLBCL cases generally have the largest mean minor axis lengths of all lymphoma subtypes. In (viii), we show a Kernel density estimation plot. The distribution of minor axis length is clearly further to the right for DLBCL than for CHL and MCL. In (ix), we show a feature importance analysis plot by SHAP values. The top 10 nuclear features by percentage importance using SHAP are shown: minor axis length interquartile range (IQR) is the second-most important feature. All of the top 10 features are area-shape features.

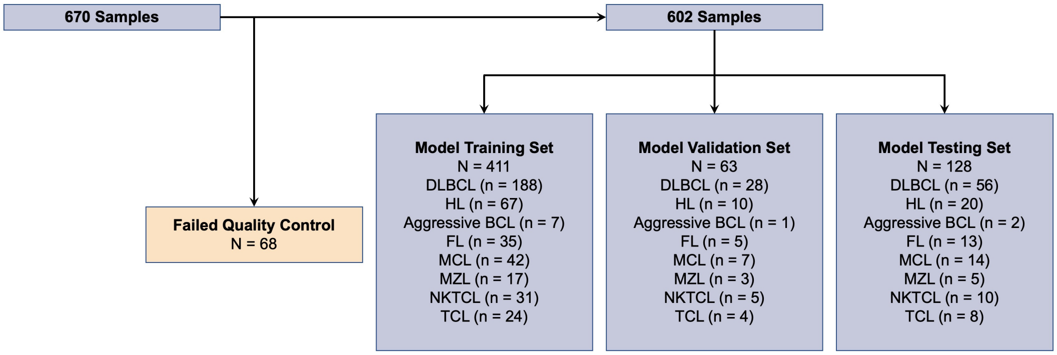


Figure A.3: Of 670 FFPE biopsy specimens, 68 failed quality control (e.g. did not have sufficient tissue per core, at least two full samples per patient, were missing ground-truth diagnoses, missing immunohistochemical stains, etc.) and were excluded from the dataset used to train and evaluate the model. The remaining 602 samples were split at a core-level into training, validation, and test splits to ensure that all extracted image patches from the same patient are in the same data split with 70% of the total tissue microarray (TMA) cores for training, 10% for validation to tune model hyperparameters, and 20% for testing. Stratified sampling was used to proportionally represent the eight diagnostic categories in each of the training, validation, and test sets.

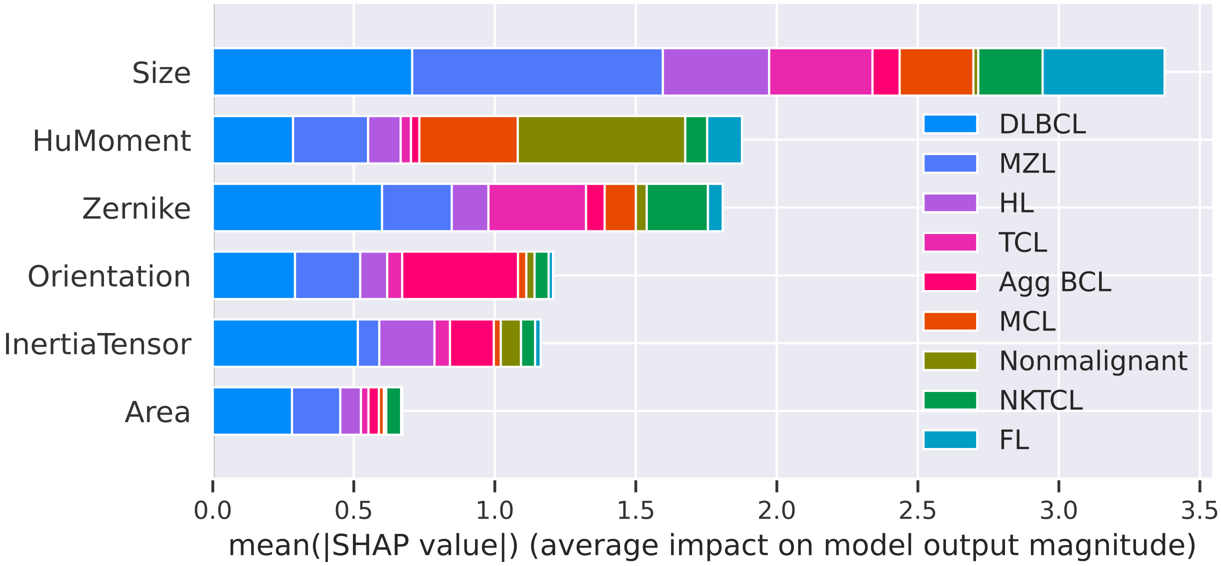


Figure A.4: Feature importance analysis by SHapley Additive exPlanation (SHAP) values, while grouping related morphological features into different categories. The nuclear size feature group has the largest mean absolute SHAP value, suggesting that of all nuclear features, size features were the most helpful for classifying DLBCL, CHL, and MCL.

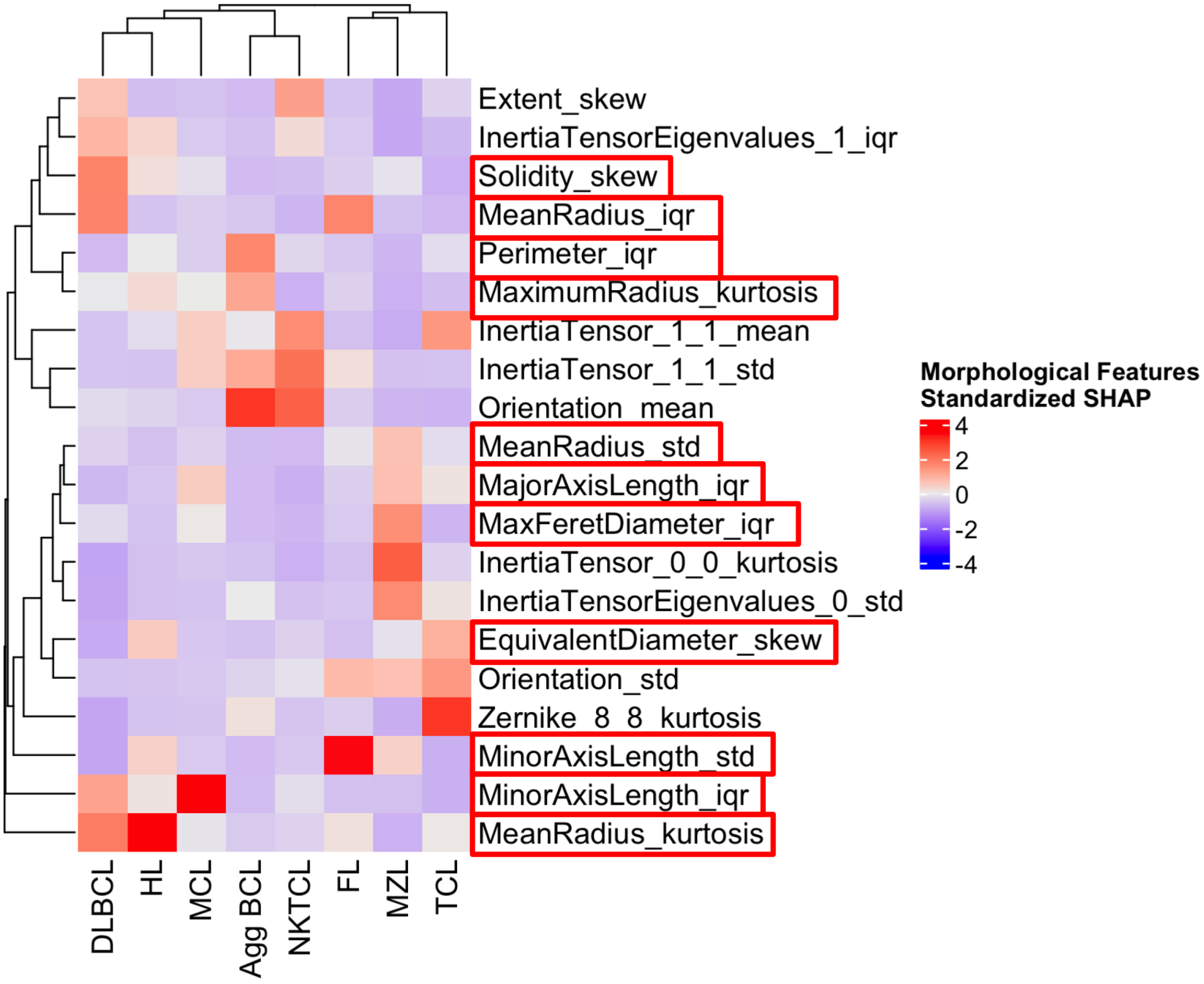


Figure A.5: The SHAP value breakdown of the top 20 morphological features in each diagnosis with the most important features in the "size" group circled.

Appendix B. Supplementary Tables

Diagnosis	Number of Cases
DLBCL	272
CHL	97
Agg BCL	10
FL	53
MCL	63
MZL	25
NKTCL	46
TCL	36

Table B.1: Dataset distribution presenting the number of available cases per lymphoma subtype.

Diagnosis	Patches (train split)	Cases (train split)
DLBCL	32926	188
CHL	11829	67
Agg BCL	1303	7
FL	6562	35
MCL	7559	42
MZL	3321	17
NKTCL	5458	31
TCL	4198	24

Table B.2: Dataset size: the number of patches/cases in the training set per-class.

Feature Name	Feature Type	Feature Description
Area	Morphological (Area Shape)	The number of pixels in the region.
Bounding Box Area	Morphological (Area Shape)	The area of a box containing the object.
Bounding Box Min/Max	Spatial / Architectural	The minimum/maximum x-, y-, and (for 3D objects) z- coordinates of the object.
Center_X, Center_Y, Center_Z	Spatial / Architectural	The x-, y-, and (for 3D objects) z- coordinates of the point farthest away from any object edge (the centroid).
Central Moment Features	Spatial / Architectural	Similar to spatial moments, but normalized to the object's centroid. These are therefore not influenced by an object's location within an image.
Compactness	Morphological (Area Shape)	The mean squared distance of the object's pixels from the centroid divided by the area. A filled circle will have a compactness of 1, with irregular objects or objects with holes having a value greater than 1.
Convex Hull Area	Morphological (Area Shape)	The number of pixels in the convex hull of the object.
Eccentricity	Morphological (Area Shape)	The eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1 (0 and 1 are degenerate cases; an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.)
Equivalent Diameter	Morphological (Area Shape)	The diameter of a circle or sphere with the same area as the object.
Euler Number	Morphological (Area Shape)	The number of objects in the region minus the number of holes in those objects, assuming 8-connectivity.
Form Factor	Morphological (Area Shape)	Calculated as $4 * \pi * \left(\frac{\text{Area}}{\text{Perimeter}}\right)^2$. Equals 1 for a perfectly circular object.
Hu Moment Features	Morphological (Area Shape)	Hu's set of image moment features. These are not altered by the object's location, size or rotation. This means that they primarily describe the shape of the object.
Inertia Tensor Features	Morphological (Area Shape)	A representation of rotational inertia of the object relative to its center.
Inertia Tensor Eigenvalues Features	Morphological (Area Shape)	Values describing the movement of the Inertia Tensor array.
Integrated Intensity	Texture / Intensity	The sum of the pixel intensities within an object.
Integrated Intensity Edge	Texture / Intensity	The sum of the edge pixel intensities of an object.
Lower Quartile Intensity	Texture / Intensity	The intensity value of the pixel for which 25% of the pixels in the object have lower values.

Feature Name	**Feature Type**	**Feature Description**
MAD Intensity	Texture / Intensity	The median absolute deviation (MAD) value of the intensities within the object. The MAD is defined as the median(| $x_i - \text{median}(x)$ |).
Major/Minor Axis Length	Morphological (Area Shape)	The length (in pixels) of the major/minor axis of the ellipse that has the same normalized second central moments as the region.
Mass Displacement	Texture / Intensity	The distance between the centers of gravity in the gray-level representation of the object and the binary representation of the object.
Max Intensity	Texture / Intensity	The maximal pixel intensity within an object.
Max Intensity Edge	Texture / Intensity	The maximal edge pixel intensity of an object.
Maximum Radius	Morphological (Area Shape)	The maximum distance of any pixel in the object to the closest pixel outside of the object. For skinny objects, this is $\frac{1}{2}$ of the maximum width of the object.
Max/Min Feret Diameter	Morphological (Area Shape)	The Feret diameter is the distance between two parallel lines tangent on either side of the object (imagine taking a caliper and measuring the object at various angles). The maximum and minimum Feret diameters are the largest and smallest possible diameters, rotating the calipers along all possible angles.
Mean Intensity	Texture / Intensity	The average pixel intensity within an object.
Mean Intensity Edge	Texture / Intensity	The average edge pixel intensity of an object.
Mean Radius	Morphological (Area Shape)	The mean distance of any pixel in the object to the closest pixel outside of the object.
Median Intensity	Texture / Intensity	The median intensity value within the object.
Median Radius	Morphological (Area Shape)	The median distance of any pixel in the object to the closest pixel outside of the object.
Min Intensity	Texture / Intensity	The minimal pixel intensity within an object.
Min Intensity Edge	Texture / Intensity	The minimal edge pixel intensity of an object.
Normalized Moment Features	Morphological (Area Shape)	Similar to central moments, but further normalized to be scale invariant. These moments are therefore not impacted by an object's size (or location).
Orientation	Morphological (Area Shape)	The angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.
Perimeter	Morphological (Area Shape)	The total number of pixels around the boundary of each region in the image.

Feature Name	Feature Type	Feature Description
Solidity	Morphological (Area Shape)	The proportion of the pixels in the convex hull that are also in the object, i.e., ObjectArea/ConvexHullArea.
Spatial Moment Features	Spatial / Architectural	A series of weighted averages representing the shape, size, rotation and location of the object.
Std Intensity	Texture / Intensity	The standard deviation of the pixel intensities within an object.
Std Intensity Edge	Texture / Intensity	The standard deviation of the edge pixel intensities of an object.
Upper Quartile Intensity	Texture / Intensity	The intensity value of the pixel for which 75% of the pixels in the object have lower values.
Zernike shape Features	Morphological (Area Shape)	These metrics of shape describe a binary object (or more precisely, a patch with background and an object in the center) in a basis of Zernike polynomials, using the coefficients as features (Boland et al., 1998). Currently, Zernike polynomials from order 0 to order 9 are calculated, giving in total 30 measurements. While there is no limit to the order which can be calculated (and indeed you could add more by adjusting the code), the higher order polynomials carry less information.

Table B.3: The full list of features extracted by our CellProfiler pipeline. All feature definitions are provided in: https://cellprofiler-manual.s3.amazonaws.com/CellProfiler-4.0.5/modules/measurement.html (Stirling et al., 2021).

Study	**Current study**	**Zhu et al. (2022)**	**Steinbuss et al. (2021)**	**Mohlman et al. (2020)**
	Large non-Hodgkin B-cell lymphoma (n = 286), Total small B-cell lymphomas (n = 154), CHL (n = 91), Total T and NK cell lymphoma (n = 96), TLL (n = 6), BLL (n = 5), Plasma cell neoplasm (n = 8), Histiocytic sarcoma (n = 1), RFH (n = 23), Total = 670	FL (n = 1129), CLL (n = 212), MCL (n = 613), DLBCL (n = 1002), CHL (n = 232), MZL (n = 888), AITL (n = 76), BL (n = 95), Total = 4247	CLL (n = 129), DLBCL (n = 119), Control LNs (n = 381), Total = 629	DLBCL (n = 36), BL (n = 34), Total = 70
Number of diagnostic categories	8	8	3	2
Average number of cases per diagnostic category	84	531	210	35
Interpretability of models	Yes	Not described	No	No
Accuracy of best model	85%	81%	95.56%	94%

Study	**Miyoshi et al. (2020)**	**Li et al. (2020)**	**Zhang et al. (2020)**
	DLBCL (n = 259), FL (n = 89), RFH (n = 40), Total = 388	DLBCL (n = 867), Non-DLBCL (n = 887), Total = 1754	CLL (n = 113), FL (n = 139), MCL (n = 122), Total = 374
Number of diagnostic categories	3	2	3
Average number of cases per diagnostic category	129	877	125
Interpret-ability of models	No	No	No
Accuracy of best model	97%	99.71-100%	99.2-100%

Study	**Achi et al. (2019)**	**Brancati et al. (2019)**	**Janowczyk and Madabhushi (2016)**
	Benign (n = 32), DLBCL (n = 32), BL (n = 32), SLL (n = 32), Total = 128	CLL (n = 113), FL (n = 139), MCL (n = 122), Total = 374	CLL (n = 113), FL (n = 139), MCL (n = 122), Total = 374
Number of diagnostic categories	4	3	3
Average number of cases per diagnostic category	32	125	125
Interpret-ability of models	No	No	No
Accuracy of best model	95%	97.06%	96.58%

Table B.4: Summary of case types assessed in studies that apply computer vision tools to lymphoma diagnosis. CLL = chronic lymphocytic leukemia, SLL = small lymphocytic lymphoma, FL = follicular lymphoma, MCL = mantle cell lymphoma, DLBCL = diffuse large B-cell lymphoma, BL = Burkitt lymphoma, RFH = reactive follicular hyperplasia, TLL = T-lymphoblastic lymphoma, BLL = B-lymphoblastic leukemia, LN = lymph nodes

Study	Current study	Yu et al. (2021)	Gupta et al. (2010)	Lesty et al. (1986)
	Large non-Hodgkin B-cell lymphoma (n = 286), Total small B-cell lymphomas (n = 154), CHL (n = 91), Total T and NK cell lymphoma (n = 96), TLL (n = 6), BLL (n = 5), Plasma cell neoplasm (n = 8), Histiocytic sarcoma (n = 1), RFH (n = 23), Total = 670	MEITL (n = 26), ITL-NOS (n = 10), Borderline cases (n = 4), Total = 40	Fine needle aspiration cytology smears of 5 cases. SLL (n = 13), FL (n = 9), DLBCL centroblastic (n = 14), DLBCL anaplastic (n = 4), DLBCL immunoblastic (n = 2), lymphoblastic lymphoma (n = 8), Total = 50	Lymphomas classified according to the International Working Formulation. Small noncleaved lymphocytes, CLL (n = 8), Predominantly small cleaved cell; follicular or diffuse or both (n = 7), Mixed diffuse small and large; cleaved or noncleaved cells (n = 13), Large diffuse noncleaved cells (n = 7), Large diffuse cleaved cells (n = 10), Total = 45
Total number of cases	670	40	50	45
Number of diagnostic categories	8	2	3*	5**
Average number of cases per diagnostic category	84	20	16	9
Interpretability of models	Yes	Yes	Yes	Yes
Accuracy of best model	85%	95%	97%	97%

Table B.5: Notable features of studies that identify interpretable features for lymphoma diagnosis. CLL = chronic lymphocytic leukemia, SLL = small lymphocytic lymphoma, FL = follicular lymphoma, MCL = mantle cell lymphoma, DLBCL = diffuse large B-cell lymphoma, BL = Burkitt lymphoma, RFH = reactive follicular hyperplasia, TLL = T-lymphoblastic lymphoma, BLL = B-lymphoblastic leukemia, WSI = whole slide images, TMA = tissue microarrays, MEITL = monomorphic epitheliotropic intestinal T-cell lymphoma, ITL-NOS = intestinal T-cell lymphoma, not otherwise specified.

*By the World Health Organization Classification, centroblastic, immunoblastic, and anaplastic variants of diffuse large B-cell lymphoma are not separate diagnostic categories.
**Diagnostic categories were based on the International Working Formulation.

Model Type	Model Features	# of features	Test Accuracy	Test Sensitivity	Test Specificity
Nuclear Morphology	Nuclear Morphological Features	310	59.7 ± 8.5	58.9 ± 10.1	84.9 ± 4.2
	Nuclear Intensity Features	225	58.9 ± 7.8	57.4 ± 9.3	84.7 ± 4.4
	Cytoplasmic Features	470	57.4 ± 9.3	55.0 ± 9.3	84.2 ± 4.4
	Nuclear Morphological + Intensity Features	630	61.2 ± 8.6	60.5 ± 10.0	87.1 ± 4.2
	Nuclear + Cytoplasmic Features	950	62.0 ± 8.5	59.7 ± 9.3	86.8 ± 3.7
	Nuclear Morphological Model: Parsimonious	36	61.2 ± 7.8	57.4 ± 9.3	84.0 ± 4.4
Architectural Features	CPArch + CT	217	44.9 ± 8.6	43.3 ± 10.6	74.7 ± 5.5
	Nuclear + CPArch	475	60.5 ± 7.7	62.2 ± 9.1	86.2 ± 4.1
	Nuclear + CPArch + CT	675	59.8 ± 7.9	52.0 ± 10.5	75.3 ± 6.2
	Nuclear + Cytoplasm + Intensity	1530	62.0 ± 8.5	62.8 ± 10.1	88.8 ± 3.3
	Nuclear + Cytoplasm + Intensity + CPArch	1595	64.3 ± 8.6	66.9 ± 6.0	88.7 ± 2.6
	Nuclear + CPArch + Cytoplasm + Intensity + CT	1695	64.1 ± 7.8	64.1 ± 8.6	86.2 ± 4.2
Deep Learning	ResNet-50 (Self-Supervised H&E)	N/A	53.5 ± 8.7	55.5 ± 10.6	82.6 ± 5.3
	TripletNet finetuned (Camelyon)	N/A	52.8 ± 8.6	53.7 ± 9.3	75.7 ± 6.0

Model Type	Model Features	Test AUROC	Test F1 Score
Nuclear Morphology	Nuclear Morphological Features	82.0 ± 6.3	54.1 ± 8.9
	Nuclear Intensity Features	82.8 ± 5.4	54.1 ± 9.9
	Cytoplasmic Features	83.1 ± 5.9	52.8 ± 9.4
	Nuclear Morphological + Intensity Features	84.2 ± 5.5	56.4 ± 9.3
	Nuclear + Cytoplasmic Features	84.6 ± 6.0	56.4 ± 9.2
	Nuclear Morphological Model: Parsimonious	81.5 ± 6.6	57.0 ± 8.6
Architectural Features	CPArch + CT	66.5 ± 7.9	39.8 ± 9.2
	Nuclear + CPArch	82.7 ± 6.7	54.7 ± 9.5
	Nuclear + CPArch + CT	76.1 ± 6.9	51.1 ± 9.4
	Nuclear + Cytoplasm + Intensity	85.3 ± 5.1	56.5 ± 9.0
	Nuclear + Cytoplasm + Intensity + CPArch	85.9 ± 2.9	58.5 ± 9.5
	Nuclear + CPArch + Cytoplasm + Intensity + CT	85.5 ± 5.0	58.0 ± 10.1
Deep Learning	ResNet-50 (Self-Supervised H&E)	82.4 ± 3.5	51.7 ± 9.8
	TripletNet finetuned (Camelyon)	81.7 ± 2.3	45.7 ± 9.6

Table B.6: Overall performance summary of feature-based models using different feature combinations and deep-learning models. All features are extracted from H&E stains only. All metrics are calculated using a weighted average across all labels. We weight by the support, the number of true instances for each label. CPArch = CellProfiler Architectural features, CT = clustering tendency.

Model Type	Model Features	Per-Class Test F1			
		DLBCL	HL	Agg BCL	FL
Nuclear Morphology	Nuclear Morphological Features	76.2 ± 8.0	65.3 ± 17.3	0.0	11.1 ± 11.1
	Nuclear Intensity Features	76.0 ± 6.6	64.0 ± 15.1	0.0	27.0 ± 25.6
	Cytoplasmic Features	77.0 ± 8.5	60.0 ± 14.4	0.0	25.0 ± 22.6
	Nuclear Morphological + Intensity Features	77.0 ± 8.0	69.4 ± 13.3	0.0	28.6 ± 25.2
	Nuclear + Cytoplasmic Features	79.0 ± 6.9	76.6 ± 13.6	0.0	34.8 ± 22.3
	Nuclear Morphological Model: Parsimonious	74.6 ± 8.0	69.2 ± 13.6	0.0	38.1 ± 25.9
Architectural Features	CPArch + CT	68.2 ± 7.8	33.3 ± 17.5	0.0	20.0 ± 20.0
	Nuclear + CPArch	76.2 ± 7.0	69.4 ± 14.6	0.0	11.1 ± 11.1
	Nuclear + CPArch + CT	75.8 ± 7.7	75.0 ± 13.4	0.0	0.0
	Nuclear + Cytoplasm + Intensity	78.1 ± 7.0	70.6 ± 14.5	0.0	30.0 ± 25.2
	Nuclear + Cytoplasm + Intensity + CPArch	78.7 ± 7.7	74.5 ± 13.4	0.0	31.6 ± 24.8
	Nuclear + CPArch + Cytoplasm + Intensity + CT	80.0 ± 7.7	70.8 ± 14.4	0.0	22.2 ± 22.2
Deep Learning	ResNet-50 (Self-Supervised H&E)	72.5 ± 8.2	59.7 ± 17.7	0.0	27.7 ± 24.5
	TripletNet finetuned (Camelyon)	70.3 ± 8.1	30.1 ± 21.0	0.0	30.6 ± 27.7

Model Type	Model Features	Per-Class Test F1			
		MCL	MZL	NKTCL	TCL
Nuclear Morphology	Nuclear Morphological Features	51.6 ± 19.4	22.2 ± 22.2	42.9 ± 29.1	0.0
	Nuclear Intensity Features	56.0 ± 20.2	0.0	29.0 ± 29.0	0.0
	Cytoplasmic Features	51.0 ± 19.6	0.0	25.0 ± 25.0	0.0
	Nuclear Morphological + Intensity Features	66.7 ± 16.6	0.0	26.7 ± 26.7	0.0
	Nuclear + Cytoplasmic Features	51.6 ± 19.4	0.0	14.3 ± 14.3	0.0
	Nuclear Morphological Model: Parsimonious	64.5 ± 17.9	25.0 ± 25.0	26.7 ± 26.7	0.0
Architectural Features	CPArch + CT	16.7 ± 16.7	33.3 ± 33.3	0.0	0.0
	Nuclear + CPArch	53.3 ± 18.9	22.2 ± 22.2	40.0 ± 30.0	0.0
	Nuclear + CPArch + CT	46.7 ± 20.0	0.0	16.7 ± 16.7	0.0
	Nuclear + Cytoplasm + Intensity	62.1 ± 17.9	0.0	25.0 ± 25.0	0.0
	Nuclear + Cytoplasm + Intensity + CPArch	71.0 ± 16.0	0.0	25.0 ± 25.0	0.0
	Nuclear + CPArch + Cytoplasm + Intensity + CT	71.0 ± 16.0	0.0	26.7 ± 26.7	0.0
Deep Learning	ResNet-50 (Self-Supervised H&E)	51.1 ± 23.9	0.0	34.1 ± 25.2	0.0
	TripletNet finetuned (Camelyon)	49.2 ± 20.6	0.0	29.0 ± 29.0	0.0

Table B.7: Per-class performance summary of feature-based models using different feature combinations and deep-learning models. All features are extracted from H&E stains only. CPArch = CellProfiler Architectural features, CT = clustering tendency.

Method	Test Accuracy	Test Weighted Sensitivity	Test Weighted Specificity	Test Weighted AUROC	F1 Score
Baseline: Hematopathologist 1 on H&E TMAs	73.0 ± 7.4	72.5 ± 8.2	73.2 ± 7.6	N/A	73.1 ± 6.6
Baseline: Hematopathologist 2 on H&E TMAs	73.0 ± 6.7	72.9 ± 8.1	73.0 ± 7.9	N/A	73.0 ± 7.0
Baseline: Hematopathologist 3 on WSIs	83.8 ± 5.4	82.8 ± 6.2	82.6 ± 6.6	N/A	83.8 ± 5.4
Baseline: General Pathologist on WSIs	73.6 ± 6.8	69.0 ± 7.6	68.3 ± 8.6	N/A	73.1 ± 6.9
Best Model with nuclear size/area features only	76.0 ± 6.9	76.0 ± 8.5	74.7 ± 9.0	81.7 ± 7.5	75.9 ± 7.1
Best H&E Model	79.8 ± 6.2	81.3 ± 7.1	81.8 ± 6.8	78.9 ± 6.9	81.5 ± 6.9

Table B.8: Performance comparison of the Best H&E Model to hematopathologists on the DLBCL vs non-DLBCL classification task.

Method	Test Accuracy	Test Weighted Sensitivity	Test Weighted Specificity	Test Weighted AUROC	Test Weighted F1 Score
Baseline: Best H&E Model	69.0 ± 7.7	69.8 ± 8.5	86.4 ± 4.5	85.5 ± 5.2	65.7 ± 8.4
Baseline: CD10, CD20, CD3, EBV ISH, BCL1, CD30	75.2 ± 7.0	75.2 ± 7.0	86.4 ± 5.4	89.7 ± 4.5	70.9 ± 8.7
Baseline: best model with all immunostains	86.1 ± 6.1	86.0 ± 7.0	93.2 ± 3.4	96.7 ± 2.2	85.1 ± 6.0
H&E + CD20	78.3 ± 6.2	79.1 ± 7.7	93.1 ± 3.5	93.3 ± 3.5	78.1 ± 6.9
H&E + CD3	79.1 ± 6.2	81.4 ± 7.0	92.9 ± 3.8	93.2 ± 3.3	78.3 ± 7.4
H&E + CD3, CD20	82.9 ± 6.2	82.9 ± 7.8	94.2 ± 3.3	94.7 ± 2.8	82.6 ± 6.1
H&E + CD3, CD20, BCL1	83.7 ± 6.2	83.7 ± 7.0	94.4 ± 3.3	95.5 ± 2.6	83.4 ± 6.1
H&E + CD10, CD20, CD3, EBV ISH, BCL1, CD30	85.3 ± 5.4	84.5 ± 7.0	93.5 ± 3.7	95.7 ± 2.7	84.7 ± 6.5

Method	Per-Class F1				
	B-Cell	HL	FL, MZL	MCL	T-cell
Baseline: Best H&E Model	78.0 ± 7.2	72.3 ± 12.8	62.9 ± 17.1	75.9 ± 15.5	17.4 ± 21.1
Baseline: CD10, CD20, CD3, EBV ISH, BCL1, CD30	78.7 ± 7.2	81.1 ± 12.2	9.1 ± 9.1	87.5 ± 10.2	87.2 ± 9.8
Baseline: best model with all immunostains	86.7 ± 6.0	92.3 ± 7.7	58.1 ± 19.3	93.3 ± 6.7	97.3 ± 2.7
H&E + CD20	85.0 ± 6.3	74.4 ± 13.4	66.7 ± 15.7	80.0 ± 13.8	74.3 ± 14.6
H&E + CD3	84.7 ± 6.5	76.2 ± 13.3	58.1 ± 18.4	75.9 ± 15.0	86.5 ± 10.3
H&E + CD3, CD20	87.7 ± 4.5	82.1 ± 11.5	66.7 ± 15.4	80.0 ± 14.1	89.5 ± 8.1
H&E + CD3, CD20, BCL1	85.7 ± 6.2	85.0 ± 10.5	64.9 ± 15.7	90.3 ± 9.7	91.9 ± 8.1
H&E + CD10, CD20, CD3, EBV ISH, BCL1, CD30	87.2 ± 5.9	84.2 ± 10.2	70.6 ± 16.1	90.3 ± 9.7	91.9 ± 8.1

Table B.9: Performance comparison of the best H&E-only model, models using only immunostains, and models using features from H&E combined with selected immunostains. All experiments are performed on the five-way grouped classification task.

Comparison		**Accuracy Difference (Method 1 - Method 2)**	**Two-tailed Paired t-test (Bootstrapping)**		**Test for Equivalence (TOST with Bootstrapping)**				**Conclusion** ($\alpha = 5\%$)
Method 1	Method 2		95% CI Lower Bound	95% CI Upper Bound	90% CI Lower Bound	$-\Delta$	90% CI Upper Bound	$+\Delta$	
Best H&E Model	Hemato-pathologist 1 on TMAs	0.082	-0.030	0.193	-0.015	-0.050	0.173	0.050	Non-inferior
Best H&E Model	Hemato-pathologist 2 on TMAs	0.040	-0.072	0.155	-0.056	-0.050	0.138	0.050	–
Best H&E Model	Hemato-pathologist 3 on WSIs	0.005	-0.119	0.119	-0.090	-0.050	0.102	0.050	–
Best H&E Model	General Pathologist on WSIs	0.080	-0.036	0.197	-0.021	-0.050	0.175	0.050	Non-inferior

Comparison		**Accuracy Difference (Method 1 - Method 2)**	**Two-tailed Paired t-test (Bootstrapping)**		**Test for Equivalence (TOST with Bootstrapping)**				**Conclusion** ($\alpha = 5\%$)
Method 1	Method 2		95% CI Lower Bound	95% CI Upper Bound	90% CI Lower Bound	$-\Delta$	90% CI Upper Bound	$+\Delta$	
Best H&E Model	TripletNet finetuned (Camelyon)	0.121	0.003	0.246	–	-0.050	–	0.050	Significant Difference

Table B.10: Summary of statistical tests comparing the best H&E-only model to pathologists and deep-learning models.

Comparison		Accuracy Difference (Method 1 - Method 2)	Two-tailed Paired t-test (Bootstrapping)		Test for Equivalence (TOST with Bootstrapping)				Conclusion ($\alpha = 5\%$)
Method 1	Method 2		95% CI Lower Bound	95% CI Upper Bound	90% CI Lower Bound	$-\Delta$	90% CI Upper Bound	$+\Delta$	
6 Stains Only (CD10, CD20, CD3, EBV ISH, BCL1, CD30)	All 46 Stains	-0.112	-0.209	-0.016	–	-0.050	–	0.050	Significant Difference
H&E Features Only		-0.173	-0.271	-0.078	–	-0.050	–	0.050	Significant Difference
H&E Features + 6 Stains		-0.009	-0.101	0.070	-0.085	-0.050	0.062	0.050	–

Table B.11: Summary of statistical tests comparing different models (model with selected immunostains, model using features extracted from H&E and selected immunostains) to a baseline model using all 46 available immunostains.

Number of Patches per Core	Patch-Level CV Accuracy	Core-Level CV Accuracy
1	54.1%	55.4%
4	52.4%	58.3%
9	48.7%	57.4%
16	46.8%	57.9%
25	44.9%	57.5%
36	43.3%	55.2%
49	42.7%	53.6%
64	41.6%	53.9%
81	40.9%	53.6%
100	40.8%	53.6%

Table B.12: There is no standard patch size (Steinbuss et al., 2021) so we performed patch-resolution experiments to select the best patch size for feature-based models (at the extreme, using one patch to represent the core). We present patch-level and core-level cross-validation (CV) accuracies for the nuclei-only model trained using different numbers of patches per-core. We considered cases when each core was divided into a perfect square number of patches (1, 4, 9, . . . , 100 patches). Using this patch extraction method, we divided the width and height of each TMA core into a fixed number of segments to produce a grid of equally-sized patches. Since TMA cores come in varying sizes, this method preserves the same label distribution in the patch-level dataset as in the original core-level dataset as we simply scale up the number of examples by a constant. We compared models fitted using features aggregated from patches of different sizes, and selected the best model based on the 5-fold cross-validation accuracy. We found experimentally that extracting a small number of patches per core (specifically, 4 patches per core) led to the best model performance, and in particular, better performance than core-level model training and prediction.