scientific data

OPEN

DATA DESCRIPTOR

Check for updates

DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set

Damir Vrabac[1,4], Akshay Smit[1,4], Rebecca Rojansky[2], Yasodha Natkunam[2], Ranjana H. Advani[3], Andrew Y. Ng[1], Sebastian Fernandez-Pol[2,4] & Pranav Rajpurkar[1,4] ✉

Diffuse Large B-Cell Lymphoma (DLBCL) is the most common non-Hodgkin lymphoma. Though histologically DLBCL shows varying morphologies, no morphologic features have been consistently demonstrated to correlate with prognosis. We present a morphologic analysis of histology sections from 209 DLBCL cases with associated clinical and cytogenetic data. Duplicate tissue core sections were arranged in tissue microarrays (TMAs), and replicate sections were stained with H&E and immunohistochemical stains for CD10, BCL6, MUM1, BCL2, and MYC. The TMAs are accompanied by pathologist-annotated regions-of-interest (ROIs) that identify areas of tissue representative of DLBCL. We used a deep learning model to segment all tumor nuclei in the ROIs, and computed several geometric features for each segmented nucleus. We fit a Cox proportional hazards model to demonstrate the utility of these geometric features in predicting survival outcome, and found that it achieved a C-index (95% CI) of 0.635 (0.574,0.691). Our finding suggests that geometric features computed from tumor nuclei are of prognostic importance, and should be validated in prospective studies.

Background & Summary

Diffuse Large B-Cell Lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma (NHL), accounting for over a third of cases[1] with more than 20,000 patients diagnosed annually in the United States[2]. DLBCL is fatal without treatment, however approximately 70% of patients can be cured with contemporary therapeutic regimens[3]. Treatment outcomes following standard R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) therapy are highly variable, and depend on a number of clinical, biologic, and genetic factors. Currently, the most effective prognostic classification is the National Comprehensive Cancer Network International Prognostic Index (NCCN-IPI), which incorporates five clinical variables including age, lactate dehydrogenase (LDH), extra-nodal sites of involvement, Ann Arbor stage, and ECOG performance status[4]. The NCCN-IPI model is widely used to risk stratify patients into good, intermediate, and poor-risk categories, however it is insufficient to guide therapeutic decision-making for individual patients.

Gene expression profiling (GEP) studies revealed distinct subtypes of DLBCL that correspond to differences in cell of origin (COO) and show different outcomes in response to R-CHOP therapy[5,6]. This approach categorizes DLBCL as either germinal center B-cell (GCB), activated B-cell (ABC), or indeterminate, based on the phase of B-cell development it most closely resembles[7]. A practical algorithm employing this approach for immunohistochemically stained, formalin fixed, paraffin embedded tissue was developed by Hans *et al.*, and despite imperfect concordance with the gold standard GEP method, it is now the most widely used algorithm in the United States for DLBCL[8]. The GCB subtype is associated with more favorable outcomes than the non-GCB subtype[5,9–13].

In addition to COO subtyping, double-hit lymphomas with concurrent chromosomal translocations of the MYC and BCL2 genes, or less commonly MYC and BCL6 genes, and double-expressor lymphomas with dual overexpression of MYC and BCL2 proteins have been found to correlate with an aggressive clinical course and poor outcomes

[1]Department of Computer Science, Stanford University, Stanford, United States. [2]Department of Pathology, Stanford University School of Medicine, Stanford, United States. [3]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford, United States. [4]These authors contributed equally: Damir Vrabac, Akshay Smit, Sebastian Fernandez-Pol, Pranav Rajpurkar. ✉e-mail: pranavsr@stanford.edu

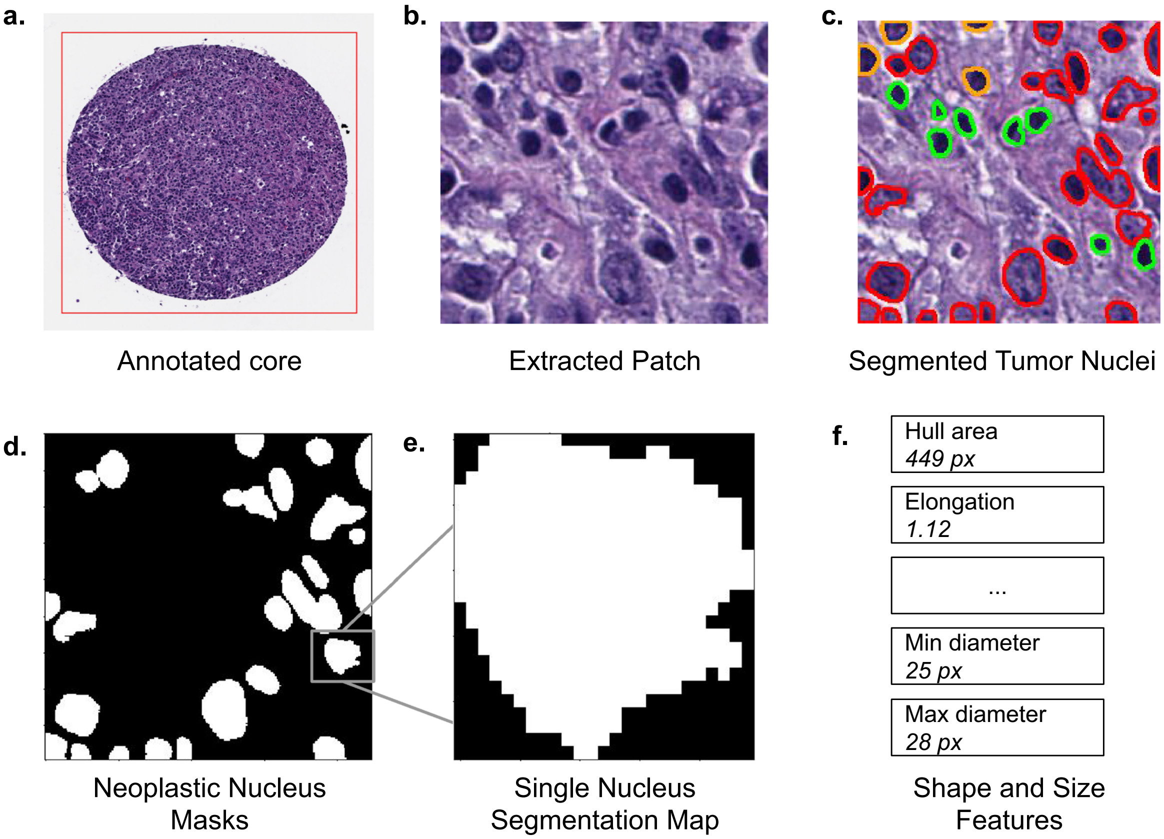

Fig. 1 Data pipeline for a single core from an H&E stained tissue microarray (TMA). In (**a**) the red rectangle is the pathologist-annotated ROI. In (**c**) red corresponds to cell nuclei classified as "neoplastic" by HoVer-Net. Green corresponds to "inflammatory" and orange corresponds to "non-neoplastic epithelial".

when treated with R-CHOP[14]. Determination of these molecular subsets is now standard of care per the World Health Organization (WHO) guidelines and patients harboring dual chromosomal translocations are now formally classified as having high grade B-cell lymphoma, with MYC and BCL2 and/or BCL6 translocations (HGBL)[15].

While COO subtyping by the Hans algorithm corresponds to morphologically distinct benign precursors, germinal center type B-cells and activated B-cells, classification based on the morphologic properties of the tumor itself has historically been challenging due to the significant histomorphologic heterogeneity of DLBCL. Cytologically, DLBCL may resemble centroblasts with multiple peripheral nucleoli and vesicular chromatin or immunoblasts with abundant cytoplasm and a single prominent nucleolus. However, the prognostic significance of these and other recognised cytologic variants, for example anaplastic type DLBCL, is unclear and the subject of continued debate[16–21].

Though several studies have thus far failed to conclusively demonstrate that morphologic classification can predict outcomes in DLBCL, automated imaging methods could potentially identify novel, prognostically significant morphological or immunohistochemical biomarkers. The ability of automated methods to identify prognostically relevant features on H&E sections that have eluded pathologists has been demonstrated[22–24]. If successful, automated image analysis could be scaled up into a cost-effective alternative to current classification methods which are typically costly and/or labor intensive. A critical requirement for the development of these models is the availability of datasets containing digitally scanned slides stained to show cell morphology and expression of relevant proteins with accompanying prognostic outcome data.

Here we present DLBCL-Morph, a publicly available dataset containing 42 digitally scanned high-resolution tissue microarrays (TMAs) from 209 DLBCL cases at Stanford Hospital. Each TMA was stained for H&E as well as for CD10, BCL6, MUM1, BCL2, and MYC protein expression. All of the TMAs are accompanied by pathologist-annotated regions-of-interest (ROIs) that indicate areas representative of DLBCL. For each patient in the dataset, we provide survival data, follow-up status, and a wide range of clinical and molecular variables such as age and MYC/BCL2/BCL6 gene translocations. We also segmented out tumor nuclei from ROIs inside the H&E stained TMAs, and provide several geometric features for each tumor nucleus.

Methods

Our dataset contains digitally scanned TMAs accompanied by pathologist-annotated ROIs. We extracted patches from the ROIs inside the H&E stained TMAs, and used a deep learning model called HoVer-Net[25] to segment tumor cell nuclei. We then computed several geometric descriptors for each segmented nucleus. Figure 1 shows our pipeline for an H&E stained TMA core. Our project was approved by the Institutional Review Board of

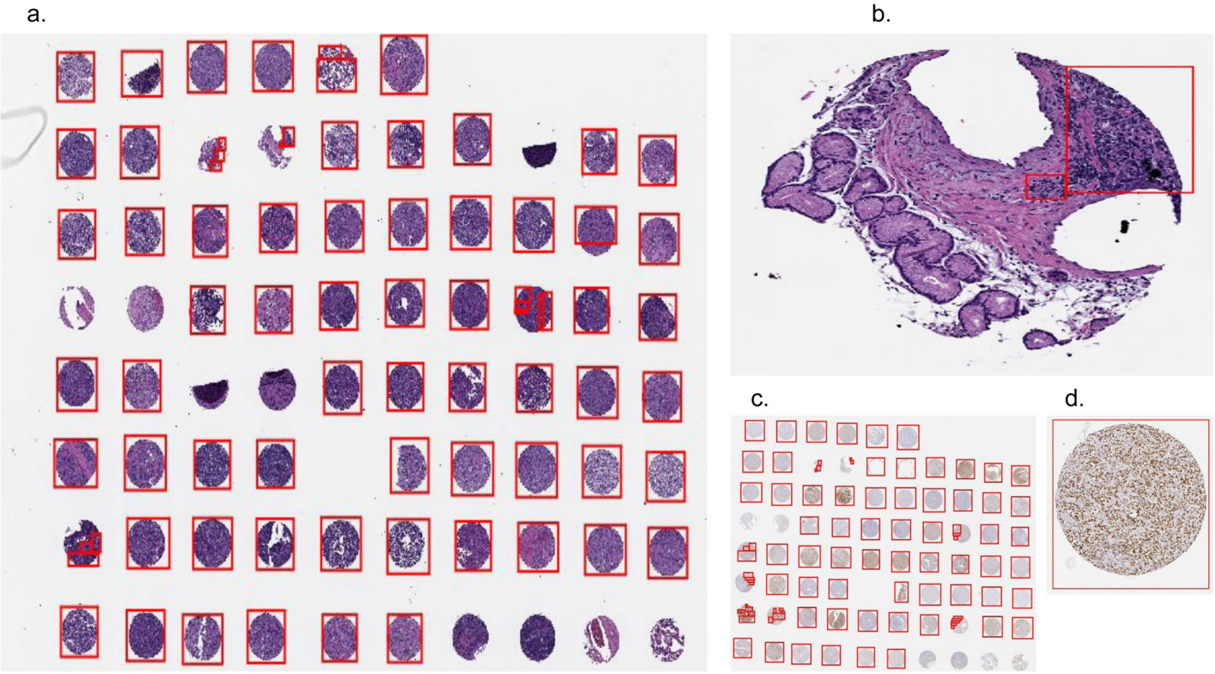

Fig. 2 Tissue microarrays (TMAs) with region-of-interest (ROI) annotations. (**a**) H&E stained TMA. The red rectangles denote ROIs annotated by a human expert. Some missing or unrepresentative cores have no ROIs. (**b**) A single core from the TMA in (**a**) with ROI that ignores unrepresentative areas of the core. (**c**) BCL6 stained TMA, containing cores from the same patients as (**a**). (**d**) A single annotated core from the TMA in (**c**). Cells stained orange show greater BCL6 expression.

Stanford University. All protected health information was removed and the project had no impact on clinical care, so the requirement for individual patient consent was waived.

Patient Cohort. The study cohort consists of patients with de novo, CD20+ DLBCL treated with curative intent with R-CHOP or R-CHOP–like immunochemotherapy with available clinical data from the Stanford Cancer Institute, Stanford, California. This patient cohort was included in a prior study with clinicopathologic inclusion criteria are as previously described[26].

Tissue Microarray. Stained tissue microarray (TMA) slides 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. This high magnification level displays the tissue in very fine detail, which we believe to be beneficial for the development of automated imaging models. Each SVS file also contains a slide label image, a macro camera image, and a thumbnail image. The slide label image is a low-resolution image of the slide's label, which shows the TMA number and the stain (eg: BCL2). The macro camera image is a low-resolution picture of the entire slide. The thumbnail is an image of the whole scanned TMA.

Our dataset includes 7 TMAs, each with a 0.4 micron thick formalin-fixed, paraffin-embedded (FFPE) section of tumors assembled in a grid. Within the microarray each tumor is represented by a 0.6-mm core diameter sample in duplicate. Replicates of each TMA were stained with H&E, which shows cell morphology. They were also stained for the expression of the following 5 oncogenes: CD10, BCL6, MUM1, BCL2, and MYC. We therefore have 6 stains per TMA, resulting in 42 distinct digitally-scanned slides. An example of an H&E stained TMA is shown in Fig. 2a and a BCL6 stained TMA is shown in Fig. 2c. Since overexpression of one or more of these proteins is observed in a significant portion of DLBCL cases, automated imaging models can use the immunostained TMAs to potentially identify prognostically significant features related to protein expression.

Pathologist annotations. Although TMA cores were already taken from areas of tissue showing DLBCL, some of the cores were partially or entirely missing. Furthermore, some cores still contained areas of tissue that had very few or no tumor cells. We obtained rectangular ROI annotations from expert pathologists to highlight the core regions which represent DLBCL accurately. The annotations were created for all TMAs and all stains at 40x magnification. The pixel coordinates for the rectangles in ROIs, along with the corresponding deidentified unique patient_id, are included in our dataset. We believe the exclusion of missing or insufficiently representative tissue areas will be beneficial for automated prognostic models which use patches from the TMAs as input. Example ROI annotations are shown in Fig. 2b,d).

Patches from stained TMAs. We extracted patches of size 224x224 from within the ROIs in the stained TMAs, at 40x magnification. The patches were extracted uniformly from inside each annotated rectangle, starting from the

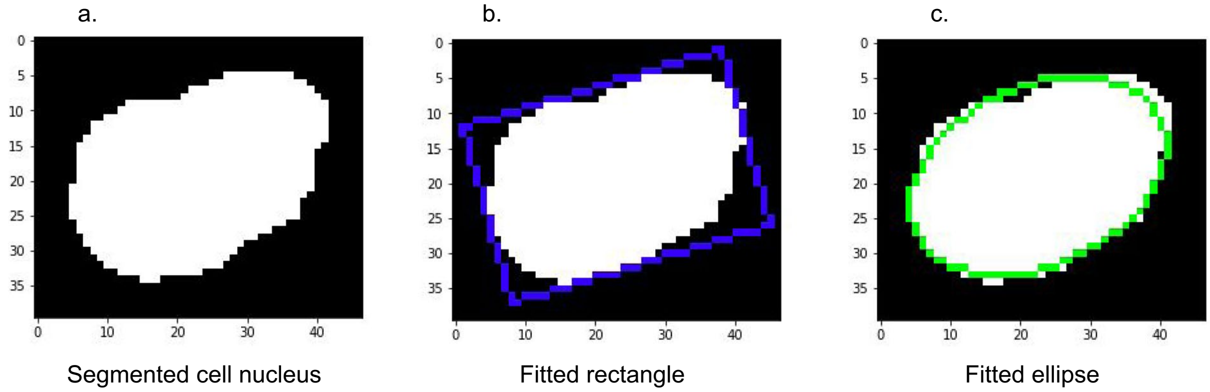

Fig. 3 Rectangle and ellipse fitted to a single segmented tumor nucleus. (**a**) a binary segmentation image for a tumor cell nucleus. For visual clarity, the image is zero-padded by 5 pixels on each side. (**b**) rotated rectangle fit to the nucleus. Our dataset provides the rectangle's center coordinates, width, height and rotation angle. (**c**) rotated ellipse fit to the nucleus. Our dataset provides the ellipse's center coordinates, perimeter, area, and major and minor axis lengths.

top-left corner and proceeding until the bottom-right corner. The patches are non-overlapping, and we omitted patches that are mostly white and contain little tissue. We provide these patches as part of our dataset. Due to our ROI annotation process detailed above, our patches exclude missing and unrepresentative areas of cores. Since deep learning based imaging methods typically cannot directly operate on images as large as the 40x magnification image, the patches can instead be used as input. We also used patches from H&E stained TMAs to segment tumor cell nuclei as described below.

Tumor cell nucleus segmentation. We used a deep learning based nucleus segmentation and classification model called HoVer-Net to segment every tumor cell inside each of the patches from H&E stained TMAs. The HoVer-Net operates independently on each patch, and produces an output image segmenting all individual cell nuclei in the patch, and another output image specifying the classification of each segmented nucleus. The HoVer-Net classifies segmented nuclei into 5 categories: neoplastic, non-neoplastic epithelial, inflammatory, connective, dead. HoVer-Net uses a neural network based on a pretrained ResNet-50 architecture to extract image features. These extracted features are then processed in three steps: the nuclear pixel (NP) step, HoVer step, and nuclear classification (NC) step. The NP step determines whether each pixel belongs to a nucleus or the background, and the HoVer step predicts the vertical and horizontal distances of nucleus pixels to their centroid, thereby allowing separation of touching nuclei. Then the NC step classifies each nucleus pixel, and aggregates these across all pixels in a segmented nucleus to classify each nucleus as neoplastic, non-neoplastic epithelial, inflammatory, connective, or dead. We used the HoVer-Net output to identify each neoplastic cell nucleus in a patch, and saved it as a separate binary image, thereby obtaining one binary image for each tumor cell. Each binary image illustrates the size and shape of the nucleus, and we provide these in our dataset. An example binary image is shown in Fig. 1e and another is shown in Fig. 3a. We used these binary images to compute geometric features for each tumor cell nucleus as described below.

Geometric features from tumor nuclei. We used the per-nucleus binary segmentation images to compute several geometric features for each tumor cell nucleus. While end-to-end imaging models may not require such hand-crafted features, prognostic models which use these features can give more explainable results, and can more clearly indicate the prognostic importance of these features.

We fit a (possibly rotated) rectangle of minimum area enclosing the binary mask, and provide the rectangle's top left point coordinates, width and height, and rotation angle. An example rectangle is shown in Fig. 3b. The rectangle's top left point is a tuple corresponding to the feature rectCenter. The first element of the tuple corresponds to the x-coordinate, and the second element corresponds to the y-coordinate. The width and height are in a tuple corresponding to the feature rectDimension. The first element of the tuple corresponds to the width, and the second element to the height. The rotation angle corresponds to the feature rotate_angle, which ranges from −90° to 0°. A value of −90° corresponds to an axis-aligned rectangle. As the rectangle is rotated clockwise, the angle increases toward 0°, at which point the rectangle is again axis-aligned and the angle resets to −90°.

We fit an ellipse around the nucleus in the binary segmentation mask, and provide the ellipse center, major axis, minor axis, perimeter and area of the ellipse. An example ellipse is shown in Fig. 3c. The ellip_centroid parameter is a tuple containing the x and y coordinates of the ellipse. The features shortAxis and longAxis correspond to the lengths of the minor and major axes respectively. The feature ellip_perimt corresponds to the ellipse perimeter, and ellip_area corresponds to the ellipse area.

We computed the maximum and minimum Feret diameters for each segmented nucleus, and provide the corresponding angles. Given an object and a fixed direction, the Feret diameter is the distance between two parallel tangents to the object, where the tangents are perpendicular to the fixed direction. The feature maxDiameter contains the Feret diameter maximized over all directions, and maxAngle specifies the angle (between −180°

and 180°) at which the maximum diameter is obtained. The features minDiameter and minAngle are similar but for the minimum Feret diameter. We further computed the convex hull of the segmented nucleus. The feature hull_area corresponds to the area of the convex hull.

Finally we computed a number of geometric features derived from the quantities described above. These features are esf, csf, sf1, sf2, elongation, and convexity. These are defined below in (1)–(6). The esf, sf1, sf2 and elongation are all simple ratios that can be thought of as measures of how “elongated” the nucleus is. In particular, they are all equal to 1 if the nucleus is perfectly circular. The csf is similar: it is a measure of circularity, and is equal to 1 if the nucleus is perfectly circular. For increasingly elliptical nuclei, the csf decreases towards 0.

$$\mathrm{esf} = \frac{\mathrm{shortAxis}}{\mathrm{longAxis}} \tag{1}$$

$$\mathrm{csf} = \frac{4\pi * \mathrm{ellip\_area}}{\mathrm{ellip\_perimt}^2} \tag{2}$$

$$\mathrm{sf1} = \frac{\mathrm{shortAxis}}{\max \mathrm{Diameter}} \tag{3}$$

$$\mathrm{sf2} = \frac{\min \mathrm{Diameter}}{\max \mathrm{Diameter}} \tag{4}$$

$$\mathrm{elongation} = \frac{\max \mathrm{Diameter}}{\min \mathrm{Diameter}} \tag{5}$$

$$\mathrm{convexity} = \sqrt{\frac{\mathrm{ellip\_area}}{\mathrm{hull\_area}}} \tag{6}$$

Data Records

The DLBCL-morph dataset[27] is organized into three folders, *TMA*, *Patches*, and *Cells* as is shown by Fig. 4. The clinical data of the patients together with the outcome is stored in *clinical_data.xlsx* and *clinical_data_cleaned.csv* where the latter contains all the patients for which the outcome is recorded and all categorical variables are converted to numerical values, e.g. ‘neg’, ‘pos’, and ‘no data’ were converted to 0, 1, and NaN, respectively for the variable CD10 IHC. Each patient has a unique identifier. There are 209 patients recorded in *clinical_data_cleaned.csv*. The column *OS* records the overall survival which is the length of time (in years) from the end of treatment until death or last follow-up. The column *Follow-up Status* (FUS) is 1 if the patient was deceased at the time of last follow-up, else 0.

TMA.

The *TMA* folder contains a total of 42 digitally-scanned TMAs, which are organized within subfolders for each stain. The filename of each TMA is a TMA id which is the same across all stains, i.e. *DLBCL-Morph/TMA/HE/TMA255* and *DLBCLMorph/TMA/BCL2/TMA255* contains cores of the same set of patients. The TMA id together with the row and column number of each core, starting with 0 and 0, respectively in the upper left corner, can be linked to the patient id through *core.csv*, each patient has two cores. The *annotations.csv* contains coordinates of ROIs annotated by human experts. For each annotation there is a patient id, TMA id, and stain where the TMA id and the stain is used to locate the TMA file that the annotation belongs to. The annotations are rectangular and the coordinates record the upper left and lower right corners based on the 40x magnification level of the TMAs.

Patches.

The Patches folder contains subfolders of stains which contains subfolders of patients that has at least one ROI. The patches are localized in the folders of patient ids with a patch id as the filename and are stored in PNG format. There are 195 patients that have at least one patch from at least one stained TMA. However, some patients do not contain patches for all 6 stains, which can occur if the core for a particular stain was missing or not covered by any ROIs.

Cells.

The *Cells* folder contains subfolders of patient ids which contains subfolders of patch ids. The binary segmentation images for tumor cell nucleus are localized in the folders of patch ids with the cell number as the filename and stored in NPY format. The NPY format is used by the Numpy package for Python to save arrays, in this case we are storing 2-dimensional arrays with binary values as segmentations of tumor cell nucleus. The cell numbers are non-consecutive since all non-tumor cells are discarded in each patch. All the geometric features computed from tumor nuclei are stored in *cell_shapes.csv* and can be linked to the nucleus segmentation images through the patch id and the cell number.

Technical Validation

We performed survival regression using the geometric and clinical features in our dataset to measure the utility of these features in predicting prognostic outcome. This analysis was performed on the 170 patients for whom patches from H&E stained TMAs were available. For each of the geometric features computed per tumor nucleus, we computed the mean and standard deviation across all nuclei for each patient. We then fit Cox Proportional

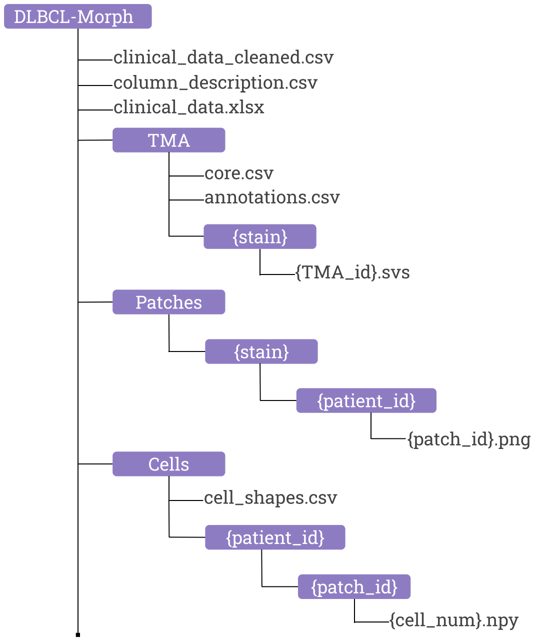

Fig. 4 The directory structure of DLBCL-Morph.

Hazards models using the binary Follow-up Status (FUS) feature as an indicator of censoring, and the overall survival (OS) feature as the time to event or censoring (in years). We evaluated our models using Harrel's C-index[28]. Random prediction would give a C-index of 0.5. Specifically we fit three models: i) using both clinical and geometric features ii) using only clinical features iii) using only geometric features.

We used the bootstrap method to obtain an "optimism-corrected" C-index[29]. We sampled 1000 bootstrap replicates with replacement and fit the model on each bootstrap replicate. We then evaluated the model on both the original data and the bootstrap replicate. We recorded the performance decrease between evaluating on the bootstrap replicate and evaluating on the original data. This decrease, averaged over all bootstrap replicates, was subtracted from the original C-index to obtain the optimism-corrected C-index. We also generated the corresponding 95% two-sided confidence intervals (CI) for the optimism-corrected C-indices using the non-parametric percentile bootstrap method[30] with 1000 bootstrap replicates.

The resulting optimism-corrected C-indices with 95% CIs for our models were: i) 0.700 (0.651,0.744) using clinical and geometric features, ii) 0.674 (0.602,0.737) using only clinical features and iii) 0.635 (0.574,0.691) using only geometric features. Thus, use of the geometric features alone allowed significantly better than random survival prediction. Use of both clinical and geometric features led to a higher performance than the use of clinical features alone, although this performance difference was not statistically significant. While prognostic classification based on the morphologic properties of the tumor has proved to be challenging and the subject of continued debate[16–21], our results suggest that geometric features computed from H&E-stained tumor nuclei can provide a significant signal to predict surival outcome. This finding should be further evaluated on external datasets and prospectively in future studies.

Usage Notes

The data is organized as shown in Fig. 4. We have provided publicly available Jupyter Notebooks[31,32] to illustrate computation of geometrical features as well as usage of the data. One notebook uses the clinical and geometric variables in the dataset to reproduce the survival regression results described in the Technical Validation section. Another notebook visualizes and reproduces the computation of several geometric features for any segmented tumor nucleus in our dataset. Finally, we provide another notebook to extract patches uniformly from inside any of the ROIs in the dataset. These patches are already included as part of the dataset, but we believe this notebook will be beneficial for researchers who work with the SVS files in our dataset. The notebooks, along with the code to compute all geometrical features from tumor nuclei as well as a link to the dataset, are provided at https://github.com/stanfordmlgroup/DLBCL-Morph.

Code availability

The code to compute all geometric features from all tumor nuclei in our dataset, along with notebooks to illustrate usage of our data and reproduce all survival regression results, is publicly available at https://github.com/stanfordmlgroup/DLBCL-Morph.

Received: 18 September 2020; Accepted: 29 March 2021;
Published online: 20 May 2021

References

1. Project, T. N.-H. L. C. A Clinical Evaluation of the International Lymphoma Study Group Classification of Non-Hodgkin's Lymphoma. *Blood* **89**, 3909–3918, https://doi.org/10.1182/blood.V89.11.3909, https://ashpublications.org/blood/articlepdf/89/11/3909/1408169/3909.pdf (1997).
2. Horvat, M. *et al.* Diffuse large b-cell lymphoma: 10 years' real-world clinical experience with rituximab plus cyclophosphamide, doxorubicin, vincristine and prednisolone. *Oncol. Lett.* https://doi.org/10.3892/ol.2018.7774 (2018).
3. Leonard, J. P. *et al.* Randomized phase II study of r-CHOP with or without bortezomib in previously untreated patients with non–germinal center b-cell–like diffuse large b-cell lymphoma. *J. Clin. Oncol.* **35**, 3538–3546, https://doi.org/10.1200/jco.2017.73.2784 (2017).
4. Zhou, Z. *et al.* An enhanced international prognostic index (NCCN-IPI) for patients with diffuse large b-cell lymphoma treated in the rituximab era. *Blood* **123**, 837–842, https://doi.org/10.1182/blood-2013-09-524108 (2014).
5. Alizadeh, A. A. *et al.* Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511, https://doi.org/10.1038/35000501 (2000).
6. Scott, D. W. Cell-of-origin in diffuse large b-cell lymphoma: Are the assays ready for the clinic? *Am. Soc. Clin. Oncol. Educ. Book* e458–e466, https://doi.org/10.14694/edbook_am.2015.35.e458 (2015).
7. Basso, K. & Dalla-Favera, R. Germinal centres and b cell lymphomagenesis. *Nat. Rev. Immunol.* **15**, 172–184, https://doi.org/10.1038/nri3814 (2015).
8. Riedell, P. A. & Smith, S. M. Should we use cell of origin and dual-protein expression in treating DLBCL? *Clin. Lymphoma Myeloma Leuk.* **18**, 91–97, https://doi.org/10.1016/j.clml.2017.12.003 (2018).
9. Gutiérrez-García, G. *et al.* Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large b-cell lymphoma treated with immunochemotherapy. *Blood* **117**, 4836–4843, https://doi.org/10.1182/blood-2010-12-322362 (2011).
10. Scott, D. W. *et al.* Prognostic significance of diffuse large b-cell lymphoma cell of origin determined by digital gene expression in formalin-fixed paraffin-embedded tissue biopsies. *J. Clin. Oncol.* **33**, 2848–2856, https://doi.org/10.1200/jco.2014.60.2383 (2015).
11. Fu, K. *et al.* Addition of rituximab to standard chemotherapy improves the survival of both the germinal center b-cell–like and non–germinal center b-cell–like subtypes of diffuse large b-cell lymphoma. *J. Clin. Oncol.* **26**, 4587–4594, https://doi.org/10.1200/jco.2007.15.9277 (2008).
12. Alizadeh, A. A. *et al.* Prediction of survival in diffuse large b-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. *Blood* **118**, 1350–1358, https://doi.org/10.1182/blood-2011-03-345272 (2011).
13. Lenz, G. *et al.* Molecular subtypes of diffuse large b-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci.* **105**, 13520–13525, https://doi.org/10.1073/pnas.0804295105 (2008).
14. Riedell, P. A. & Smith, S. M. Double hit and double expressors in lymphoma: Definition and treatment. *Cancer* **124**, 4622–4632, https://doi.org/10.1002/cncr.31646 (2018).
15. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* 127, 2375–2390, https://doi.org/10.1182/blood-2016-01-643569, https://ashpublications.org/blood/article-pdf/127/20/2375/1393632/2375.pdf (2016).
16. Engelhard, M. *et al.* Subclassification of Diffuse Large B-Cell Lymphomas According to the Kiel Classification: Distinction of Centroblastic and Immunoblastic Lymphomas Is a Significant Prognostic Risk Factor. *Blood* **89**, 2291–2297, https://doi.org/10.1182/blood.V89.7.2291, https://ashpublications.org/blood/article-pdf/89/7/2291/1642341/2291.pdf (1997).
17. Baars, J. W. *et al.* Diffuse large b-cell non-hodgkin lymphomas: the clinical relevance of histological subclassification. *Br. J. Cancer* **79**, 1770–1776, https://doi.org/10.1038/sj.bjc.6690282 (1999).
18. Diebold, J. *et al.* Diffuse large b-cell lymphoma: A clinicopathologic analysis of 444 cases classified according to the updated kiel classification. *Leuk. & Lymphoma* **43**, 97–104, https://doi.org/10.1080/10428190210173 (2002).
19. Nakamine, H. *et al.* Prognostic significance of clinical and pathologic features in diffuse large b-cell lymphoma. *Cancer* **71**, 3130–3137, 10.1002/1097-0142(19930515)71:10<3130::aid-cncr2820711039>3.0.co;2-r (1993).
20. Salar, A. *et al.* Diffuse large b-cell lymphoma: is morphologic subdivision useful in clinical management? *Eur. J. Haematol.* **60**, 202–208, https://doi.org/10.1111/j.1600-0609.1998.tb01023.x (2009).
21. Villela, L. *et al.* Prognostic features and outcome in patients with diffuse large b-cell lymphoma who do not achieve a complete response to first-line regimens. *Cancer* 91, 1557–1562, 10.1002/1097-0142(20010415)91:8<1557::aid-cncr1165>3.0.co;2-4 (2001).
22. Beck, A. H. *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Medicine* **3**, 108ra113–108ra113, https://doi.org/10.1126/scitranslmed.3002564, https://stm.sciencemag.org/content/3/108/108ra113.full.pdf (2011).
23. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Medicine* **25**, 1054–1056, https://doi.org/10.1038/s41591-019-0462-y (2019).
24. Jain, M. S. & Massoud, T. F. Predicting tumour mutational burden from histopathological images using multiscale deep learning. *bioRxiv* https://doi.org/10.1101/2020.06.15.153379, https://www.biorxiv.org/content/early/2020/06/16/2020.06.15.153379.full.pdf (2020).
25. Graham, S. *et al.* Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, 1812.06499 (2018).
26. Rosenwald, A. *et al.* Prognostic significance of myc rearrangement and translocation partner in diffuse large b-cell lymphoma: A study by the lunenburg lymphoma biomarker consortium. *J. Clin. Oncol.* **37**, 3359–3368, https://doi.org/10.1200/JCO.19.00743, PMID: 31498031 (2019).
27. Vrabac, D. *et al* Morphological feature annotation using deep learning for a clinically, histologically, and cytogenetically annotated digital image set for dlbcl. *Figshare* https://doi.org/10.6084/m9.figshare.c.5109221 (2021).
28. Harrell, J. *et al* Evaluating the Yield of Medical Tests. *JAMA* 247, 2543–2546, https://doi.org/10.1001/jama.1982.03320430047030 https://jamanetwork.com/journals/jama/articlepdf/372568/jama_247_18_030.pdf (1982).
29. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Medicine* **15**, 361–387 10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4 (1996).
30. Efron, B. & Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat. Sci.* **1**, 54–75, https://doi.org/10.1214/ss/1177013815 Publisher: Institute of Mathematical Statistics (1986).
31. Perez, F. & Granger, B. E. Ipython: A system for interactive scientific computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
32. Kluyver, T. *et al.* Jupyter notebooks -a publishing format for reproducible computational workflows. In Loizides, F. & Scmidt, B. (eds.) *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90 (IOS Press, 2016).

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-021-00915-w.

Correspondence and requests for materials should be addressed to P.R.