SUPPLEMENTARY MATERIALS FOR
“SAFARI: SHAPE ANALYSIS FOR AI-SEGMENTED IMAGES”

By Esteban Fernández Morales¹, Shengjie Yang², Sy Han Chiou¹, Chul Moon³, Cong Zhang¹, Bo Yao², Guanghua Xiao²†, and Qiwei Li¹,*

¹Department of Mathematical Sciences, The University of Texas at Dallas, *qiwei.li@utdallas.edu
²Quantitative Biology Research Center, Department of Population and Data Sciences, The University of Texas Southwestern Medical Center, †guanghua.xiao@utsouthwestern.edu
³Department of Statistical Science, Southern Methodist University
S1. Supplement to Shape Representations.

S1.1. Binary Matrix. Let $M_{W \times L}$ be a binary matrix representing an arbitrary $W$-by-$L$ image, containing a 4-connected region, where the foreground and background are composed of ones and zeros, respectively. Additionally, we can represent each pixel in the image as a point in a 2-dimensional discrete plane, that is, each entry $M_{wl} \in M$ can be denoted as a point $(l, w) \in \mathbb{N}^2$. Furthermore, to differentiate between foreground and background points, let $I_R: \mathbb{N}^2 \rightarrow \{0, 1\}$ be the indicator function for an image matrix given by

$$I_R(l, w) = \begin{cases} 1 & \text{if } (l, w) \text{ is a foreground pixel}, \\ 0 & \text{if } (l, w) \text{ is a background pixel}. \end{cases}$$

The indicator function $I_R$ and distribution of points $(l, w)$’s will be used to recreate the region’s contour in a two dimensional Cartesian plane, known as the polygonal chain.

S1.2. Polygonal Chain. The entries of the binary matrix $M_{W \times L}$ that make up the contour of the region can be extracted by the Moore-Neighbor tracing algorithm, modified by Jacob’s stopping criteria, with the 1) starting boundary point, 2) direction to traverse the boundary (clockwise or counter-clockwise), and 3) pixel connectivity (Gonzalez, Woods and Eddins, 2020). EBImage, and as a result also SAFARI, uses a 4-connectivity. Therefore, the first argument is trivial. For the starting boundary point, the point in the lowest left-most location is chosen. Specifically, let $S = \{(l_i, w_i) \mid I_R(l_i, w_i) = 1 \land w_i = \min_{1 \leq k \leq W} w_k\}$ be the collection of points that make up the region and are located in the lowest $y$-coordinate such that

$$S = \begin{cases} S_1 & |S| = 1 \\ \min S & \text{otherwise} \end{cases}$$

Applying the Moore-Neighbor tracing algorithm to the region matrix, results in the points that make up the boundary of the region, specifically, where the boundary begins at the lowest left-most area of the region and traverse through the boundary in a clockwise direction. From the boundary points, we can create a sequence of points, known as the closed polygonal chain, that represents the boundary of the region by creating a closed and simple polygon. We can see the binary matrix and its corresponding polygonal chain in Figure S2. Let $N = \sum_{l, w} I_R(l, w)$ be the total number of points that make up the region and $P_{(n+1) \times 2}$ be the collection of points that make up the closed polygonal chain of the region such that 1) $n \leq N$ is the number of boundary points, 2) $P_i = (x_i, y_i) \in \mathbb{N}^2$, for $i = 1, \ldots, n+1$, and 3) $(x_1, y_1) = (x_{n+1}, y_{n+1})$. Through the polygonal chain, we can derive further two- and one-dimensional shape representations that can be used to compute specific descriptors.

S1.3. Chain Code. The slope of a shape’s contour can be approximated by the directional changes between two consecutive boundary points. These directional changes can be encoded to, essentially, assign a number (from 0 to 7) to each possible relative direction resulting in an encoding list, each element known as the chain code, that provides a compact representation of the shape’s contour (Wirth, 2004; Agu, 2014). Let $c$ be a $1 \times n$ vector representing the chain codes of the polygonal chain where each
entry $c_i \in \mathbb{N} \cap [0, 7]$ corresponds to a direction in the 8-way split of the unit circle and determined by a series of steps.

First, we determine the angle between the vector composed of the difference between the two consecutive points and the x-axis, that is, let $\theta_i = \arctan2(d_i)$ be the resulting angle where $d_i = P_i - P_{i+1}$ is the difference. Since $\theta_i \in [-\pi, \pi]$, we have to transform the angle to

$$\hat{\theta}_i = \begin{cases} 
\theta_i & \theta_i \geq 0, \\
\theta_i + 2\pi & \theta_i < 0,
\end{cases}$$

such that $\hat{\theta}_i \in [0, 2\pi)$. As a result, we can now determine the corresponding chain code

$$c_i = \left\lfloor \frac{\hat{\theta}_i}{\pi/4} \right\rfloor.$$

Clearly, we can see that this procedure splits the unit circle into eight equal parts. Additionally, if the directional change does not exactly align within the eight splits, then the rounding operator $\lfloor \cdot \rfloor$ will approximate the chain code to the nearest integer.

S1.4. Curvature Chain Code. Let $\Delta c$ be a $1 \times n$ vector representing the curvature chain code such that each entry is formed from a transformation of the difference between two consecutive chain codes, that is, let $\Delta d_i = c_i - c_{i+1}$ and

$$\Delta c_i = \begin{cases} 
\Delta d_i - 7 & \Delta d_i > 2, \\
\Delta d_i + 7 & \Delta d_i < 2, \\
\Delta d_i & \text{otherwise.}
\end{cases}$$

This simple chain code derivation estimates the curvature and contains information on the convexity of the shape (Wirth, 2004).

S1.5. Radial Lengths. Let $r$ be a $1 \times n$ vector of radial lengths, that is, each entry $r_i = \|P_i - P_c\|$, for $i = 1, \ldots, n$, is the Euclidean distance from the boundary point to the shape’s centroid. We define the centroid of the polygonal chain $P_c = (x_c, y_c)$ as

$$x_c = \frac{1}{6A} \sum_{i=1}^{n} (x_i + x_{i+1})(y_{i+1}y_i - x_{i+1}y_i)$$

$$y_c = \frac{1}{6A} \sum_{i=1}^{n} (y_i + y_{i+1})(x_{i+1}y_i - x_{i+1}y_i)$$

where

$$A = \frac{1}{2} \sum_{i=1}^{n} (x_iy_{i+1} - x_{i+1}y_i)$$

is the signed area of the shape, obtained using Gauss’s area formula. Clearly, the radial lengths are not scale-invariant (as the Euclidean distance is not). Therefore, to properly analyze the structure of $r$, the individual radial lengths must be normalized.

S1.6. Normalized Radial Lengths. Let $r_{(n)}$ be the maximum radial length in $r$ such that we can introduce a $1 \times n$ vector of normalized radial lengths, denoted as $\tilde{r}$, where each entry $\tilde{r}_i = r_i / r_{(n)}$, $i = 1, \ldots, n$. By normalizing the radial lengths, we have obtain a 1-dimensional signal that is scale-invariant and which we can use to analyze the fine details of the shape’s contour (Wirth, 2004).
Table S1
Components of the output list resulting from the SAFARI procedure.

| Component | Description |
|-----------|-------------|
| desc      | A data.frame object of the shape features corresponding to each segmented ROI. |
| holes     | An integer matrix containing the holes within each ROI, labeled according to the regions. |
| id        | A character vector that is identical to the id argument. |
| k         | A specified factor to enlarge the polygonal chain by with the default being 3. |
| n         | Number of resulting segmented regions. |
| plg.chains| A list object where each component is the polygonal chain of a segmented ROI. |
| regions   | An integer matrix containing the segmented ROI, labeled from largest to smallest. |

*It takes about 0.3 seconds for SAFARI to run a moderate size binary image (300 \times 300 pixels).

Table S2
Data notation for shape representations.

| Name                        | Data                      | Support                  |
|-----------------------------|---------------------------|--------------------------|
| Binary Matrix               | $M_{W \times L}$          | $M_{wl} \in \{0,1\}$    |
| Polygonal Chain             | $P_{(n+1) \times 2}$      | $P_{i} = (x_{i}, y_{i}) \in \mathbb{N}^2$ |
| Convex Hull Chain           | $P_{CH}$                  | $P_{CH_{i}} \in \mathbb{P}$ |
| Minimum Bounding Box Chain  | $P_{MBB}$                 | $P_{MBB_{i}} \in \mathbb{R}_{2}^n$ |
| Chain Code                  | $c = [c_{i}]_{1 \times n}$| $c_{i} \in \mathbb{N} \cap [0,7]$ |
| Curvature Chain Code        | $\Delta c = [\Delta c_{i}]_{1 \times n}$ | $\Delta c_{i} \in \mathbb{Z} \cap [-2,2]$ |
| Radial Lengths              | $r = [r_{i}]_{1 \times n}$ | $r_{i} \in \mathbb{R}_{+}$ |
| Normalized Radial Lengths   | $\tilde{r} = [\tilde{r}_{i}]_{1 \times n}$ | $\tilde{r}_{i} \in [0,1]$ |
| Category | Data Name | Formula | Rationale | Reference |
|----------|-----------|---------|-----------|-----------|
| **Net Area** | Area(perimeter) = \( \sum_{i=1}^{n} A_i(M) \) | \( R \): Number of pixels that are in the region. | \( \star \) | \( \star \) | \( \star \) | \( \star \) |
| **Thicknes** | \( \bar{M} = \frac{A}{W} \) | **C2** | **O** | **C** | | | |
| **Planimetry** | Planimeter(\( P \)) = \( \sum_{i=1}^{n} \left( R_i \right) \) | \( R \): Number of pixels that are in the region. | \( \star \) | \( \star \) | | | |
| **Edge Area** | \( \bar{M} = \frac{A}{W} \) | **C2** | **O** | **C** | | | |
| **Perimeter** | \( P = \sum_{i=1}^{n} R_i \) | \( R \): Number of pixels that are in the region. | \( \star \) | \( \star \) | | | |
| **Circularity** | \( C = \frac{\pi P^2}{A} \) | | | | | | |
| **Ellipticality** | \( E = \frac{P}{P_{MBB}} \) | | | | | | |
| **Polar Width** | \( P = \sum_{i=1}^{n} R_i \) | \( R \): Number of pixels that are in the region. | \( \star \) | \( \star \) | | | |
| **Curvature** | Measurement can be checked by applying the same formula as in Filled Area to \( P_{FB} \) | | | | | | |
| **Geometric** | **Perimeter** | Measurement can be checked by applying the same formula as in Perimeter to \( P_{FB} \) | | | | | |
| **Roundness** | \( R = \frac{A_{MBB}}{P_{MBB}^2} \) | \( R \): Number of pixels that are in the region. | | | | | |
| **Roundness** | \( R = \frac{A_{MBB}}{P_{MBB}^2} \) | \( R \): Number of pixels that are in the region. | | | | | |
| **Major Axis Length** | \( A_{MBB}(\bar{M}) = \frac{\pi}{4} \left( R_i \right) \) | \( R \): Number of pixels that are in the region. | | | | | |
| **Major Axis Length** | \( A_{MBB}(\bar{M}) = \frac{\pi}{4} \left( R_i \right) \) | \( R \): Number of pixels that are in the region. | | | | | |
| **Bounding Box Area** | \( A_{MBB}(\bar{M}) = \frac{\pi}{4} \left( R_i \right) \) | \( R \): Number of pixels that are in the region. | | | | | |
| **Isometric** | \( E_{\text{iso}}(\bar{M}) = \frac{\pi}{2} \left( R_i \right) \) | | | | | | |
| **Nidol Area** | \( \bar{N} = \frac{\pi}{2} \left( R_i \right) \) | | | | | | |
| **Zero Crossing Contour** | \( E_{\text{cr}}(\bar{M}) = \frac{\pi}{2} \left( R_i \right) \) | | | | | | |
| **Normalized Moment Centroid** | \( S_{\text{cr}}(\bar{M}) = \frac{\pi}{2} \left( R_i \right) \) | | | | | | |
| **Typological** | \( N_{\text{id}}(\bar{M}) = \frac{\pi}{2} \left( R_i \right) \) | | | | | | |

1. We note that \( E = \left( M_{x_{1}} \right) \cdots M_{x_{n}} = M_{x_{1}} \cdots M_{x_{n}}(M_{x_{1}} - M_{x_{n}}) \) in the set containing the repeated mean steps applied to the group, where \( E \) is a structuring element.
2. The number of mean steps needed for the desired median steps.
3. The measurement applied to current region, while no set.
4. The measurement applied to current region, while no set.
5. The measurement applied to current region, while no set.
6. The measurement applied to current region, while no set.
7. The measurement applied to current region, while no set.
8. The measurement applied to current region, while no set.
9. The measurement applied to current region, while no set.
10. The measurement applied to current region, while no set.
11. The measurement applied to current region, while no set.
12. The measurement applied to current region, while no set.
13. The measurement applied to current region, while no set.
14. The measurement applied to current region, while no set.
15. The measurement applied to current region, while no set.
16. The measurement applied to current region, while no set.
17. The measurement applied to current region, while no set.
18. The measurement applied to current region, while no set.
19. The measurement applied to current region, while no set.
20. The measurement applied to current region, while no set.
21. The measurement applied to current region, while no set.
22. The measurement applied to current region, while no set.
23. The measurement applied to current region, while no set.
24. The measurement applied to current region, while no set.
25. The measurement applied to current region, while no set.
26. The measurement applied to current region, while no set.
27. The measurement applied to current region, while no set.
28. The measurement applied to current region, while no set.
29. The measurement applied to current region, while no set.
30. The measurement applied to current region, while no set.
31. The measurement applied to current region, while no set.
32. The measurement applied to current region, while no set.
33. The measurement applied to current region, while no set.
34. The measurement applied to current region, while no set.
35. The measurement applied to current region, while no set.
36. The measurement applied to current region, while no set.
37. The measurement applied to current region, while no set.
38. The measurement applied to current region, while no set.
39. The measurement applied to current region, while no set.
40. The measurement applied to current region, while no set.
41. The measurement applied to current region, while no set.
42. The measurement applied to current region, while no set.
43. The measurement applied to current region, while no set.
44. The measurement applied to current region, while no set.
45. The measurement applied to current region, while no set.
46. The measurement applied to current region, while no set.
47. The measurement applied to current region, while no set.
Table S4
Patient characteristics of the National Lung Screening Trial (NLST) and The Cancer Genome Atlas (TCGA) datasets. Values are either mean ± standard deviation, or number (percentage). In the case of the survival time, we use the median instead of the mean.

|                        | NLST       | TCGA       |
|------------------------|------------|------------|
| Number of Patients     | 143        | 61         |
| Age (in years)         | 64.01 ± 5.19 | 58.26 ± 12.48 |
| Survival Time (in days)| 1517 ± 730.04 | 403.93 ± 291.52 |
| Karnofsky Score (0-100)| –          | 81.15 ± 12.92 |
| Status                 |            |            |
| Alive                  | 98 (68.53%)| 21 (34.42%)|
| Dead                   | 45 (31.47%)| 40 (65.57%)|
| Gender                 |            |            |
| Male                   | 80 (55.94%)| 42 (68.85%)|
| Female                 | 63 (44.06%)| 19 (31.15%)|
| Smoking Status         |            |            |
| Yes                    | 75 (52.45%)| –          |
| No                     | 68 (47.55%)| –          |
| Cancer Stage           |            |            |
| Stage I                | 95 (66.43%)| –          |
| Stage II               | 15 (10.49%)| –          |
| Stage III              | 23 (16.08%)| –          |
| Stage IV               | 10 (6.99%) | –          |
Table S5
Univariate analysis of individual shape features in the National Lung Screening Trial (NLST) dataset. A Cox proportional-hazards (CoxPH) model was fitted to each centered and scaled feature, clustered to adjust for patients with multiple samples.

| Feature         | Coefficients | Hazard Ratio (HR) | Standard Error (SE) | Robust Standard Error | p-value* |
|-----------------|--------------|-------------------|---------------------|-----------------------|----------|
| Net Area        | 0.2679       | 1.3072            | 0.0864              | 0.1173                | 0.0224   |
| Thickness       | 0.2733       | 1.3143            | 0.0943              | 0.1119                | 0.0146   |
| Elongation      | -0.1467      | 0.8636            | 0.1153              | 0.1035                | 0.1561   |
| Area Filled     | 0.2675       | 1.3066            | 0.0869              | 0.1126                | 0.0175   |
| Perimeter       | 0.2900       | 1.3364            | 0.1026              | 0.1280                | 0.0235   |
| Circularity     | 0.1623       | 1.1762            | 0.1066              | 0.1163                | 0.1627   |
| Convex Area     | 0.2853       | 1.3301            | 0.0888              | 0.1182                | 0.0158   |
| Convex Perimeter| 0.3467       | 1.4144            | 0.1017              | 0.1331                | 0.0092   |
| Roundness       | 0.1249       | 1.1331            | 0.1140              | 0.1208                | 0.3243   |
| Convexity       | 0.1165       | 1.1236            | 0.1154              | 0.1359                | 0.3913   |
| Solidity        | 0.2796       | 1.3226            | 0.1210              | 0.1383                | 0.0433   |
| Major Axis Length| 0.3934    | 1.4820            | 0.1046              | 0.1359                | 0.0038   |
| Major Axis Angle  | -0.0213       | 0.9790            | 0.1114              | 0.1166                | 0.8553   |
| Minor Axis Length  | 0.2857       | 1.3307            | 0.1020              | 0.1206                | 0.0179   |
| Bounding Box Area | 0.3092       | 1.3624            | 0.0908              | 0.1175                | 0.0085   |
| Eccentricity   | -0.1456      | 0.8645            | 0.1101              | 0.1239                | 0.2398   |
| Fibre Length   | 0.3093       | 1.3625            | 0.0952              | 0.1132                | 0.0063   |
| Fibre Width    | 0.2860       | 1.3310            | 0.1028              | 0.1281                | 0.0255   |
| Curl            | -0.0705      | 0.9319            | 0.1160              | 0.1222                | 0.5649   |
| Bending Energy | 0.0054       | 1.0054            | 0.1074              | 0.1010                | 0.9572   |
| Total Abs. Curvature | 0.0074 | 1.0074            | 0.1072              | 0.0991                | 0.9405   |
| Radial Mean    | 0.0682       | 1.0706            | 0.1107              | 0.1271                | 0.5916   |
| Radial S.D.    | -0.0586      | 0.9431            | 0.1079              | 0.1052                | 0.5776   |
| Entropy        | -0.0624      | 0.9395            | 0.1060              | 0.1037                | 0.5475   |
| Area Ratio     | -0.0504      | 0.9509            | 0.1094              | 0.1131                | 0.6560   |
| Zero Crossing  | -0.0354      | 0.9652            | 0.1087              | 0.1074                | 0.7418   |
| Normalized Moment | -0.1409   | 0.8686            | 0.1152              | 0.1383                | 0.3082   |
| Number of Holes| 0.2289       | 1.2572            | 0.0844              | 0.0911                | 0.0120   |
| Number of Protrusions | 0.2877 | 1.3334            | 0.1029              | 0.1282                | 0.0248   |

*Bolding signifies features with p-value ≤ 0.05
Table S6
Comparison of the results obtained in our univariate study, compared to those in Wang et al. (2018).
The *p*-value in Wang et al. (2018) corresponds to either the sum of the shape feature for all regions or
the shape feature for the main region, whichever was more significant. Additionally, bolding signifies
features with *p*-value ≤ 0.05 and we show features not present in either study.

| Feature                     | Our Study | Wang’s Paper |
|-----------------------------|-----------|--------------|
| Average Tumor Probability   | –         | 0.78         |
| Net Area                    | 0.0224    | 0.0033       |
| Thickness                   | 0.0146    | –            |
| Elongation                  | 0.1563    | –            |
| Area Filled                 | 0.0175    | 0.0029       |
| Perimeter                   | 0.0235    | 0.0034       |
| Circularity                 | 0.1627    | 0.019*       |
| Convex Area                 | 0.0158    | 0.0047       |
| Convex Perimeter            | 0.0092    | –            |
| Roundness                   | 0.3243    | –            |
| Convexity                   | 0.3913    | –            |
| Solidity                    | 0.0433    | 0.16         |
| Major Axis Length           | 0.0038    | 0.0099       |
| Major Axis Angle            | 0.8553    | 0.92         |
| Minor Axis Length           | 0.0179    | 0.030        |
| Bounding Box Area           | 0.0085    | –            |
| Eccentricity                | 0.2398    | 0.13         |
| Extent                      | –         | 0.34         |
| Fibre Length                | 0.0063    | –            |
| Fibre Width                 | 0.0255    | –            |
| Curl                        | 0.5640    | –            |
| Bending Energy              | 0.9572    | –            |
| Total Abs. Curvature        | 0.9405    | –            |
| Radial Mean                 | 0.5916    | –            |
| Radial S.D.                 | 0.5776    | –            |
| Entropy                     | 0.5475    | –            |
| Area Ratio                  | 0.6560    | –            |
| Zero Crossing               | 0.7418    | –            |
| Normalized Moment           | 0.3082    | –            |
| Number of Holes             | 0.0120    | 0.0033       |
| Number of Protrusions       | 0.0248    | –            |
| Number of Regions           | –         | 0.48         |

*Measure corresponds to a variation of the formula we used.*
### Table S7

Univariate analysis of individual shape features in The Cancer Genome Atlas (TCGA) dataset. A Cox proportional-hazards (CoxPH) model was fitted to each centered and scaled feature.

| Feature             | Coefficients | Hazard Ratio (HR) | Standard Error (SE) | p-value  |
|---------------------|--------------|-------------------|---------------------|----------|
| Net Area            | 0.5784       | 1.7832            | 0.1706              | <0.001   |
| Thickness           | 0.327        | 1.3868            | 0.1821              | 0.0726   |
| Elongation          | -0.1706      | 0.8431            | 0.287               | 0.5522   |
| Area Filled         | 0.4361       | 1.5467            | 0.1843              | 0.0180   |
| Perimeter           | 0.4182       | 1.5192            | 0.1975              | 0.0342   |
| Circularity         | 0.0962       | 1.101             | 0.1631              | 0.5555   |
| Convex Area         | 0.555        | 1.742             | 0.1936              | 0.0041   |
| Convex Perimeter    | 0.4917       | 1.6351            | 0.2001              | 0.0140   |
| Roundness           | 0.2225       | 1.2492            | 0.1839              | 0.2264   |
| Convexity           | 0.099        | 1.1041            | 0.1743              | 0.5700   |
| Solidity            | 0.1957       | 1.2161            | 0.1807              | 0.2790   |
| Major Axis Length   | 0.4776       | 1.6121            | 0.2006              | 0.0173   |
| Major Axis Angle    | 0.5612       | 1.7528            | 0.2104              | 0.0076   |
| Minor Axis Length   | 0.5108       | 1.6666            | 0.198               | 0.0099   |
| Bounding Box Area   | 0.5371       | 1.7111            | 0.1913              | 0.005    |
| Eccentricity        | 0.0695       | 1.072             | 0.1501              | 0.6435   |
| Fibre Length        | 0.2615       | 1.2989            | 0.1776              | 0.1410   |
| Fibre Width         | 0.3954       | 1.485             | 0.1966              | 0.0443   |
| Curl                | -0.0652      | 0.9368            | 0.2386              | 0.7845   |
| Bending Energy      | 0.14         | 1.1502            | 0.165               | 0.3964   |
| Total Abs Curvature | 0.1501       | 1.1619            | 0.1655              | 0.3646   |
| Radial Mean         | 0.2061       | 1.2289            | 0.171               | 0.2281   |
| Radial Sd           | 0.0786       | 1.0818            | 0.1649              | 0.6336   |
| Entropy             | 0.146        | 1.1572            | 0.1624              | 0.3689   |
| Area Ratio          | 0.0869       | 1.0908            | 0.1696              | 0.6084   |
| Zero Crossing       | 0.2271       | 1.255             | 0.1638              | 0.1656   |
| Normalized Moment   | 0.2168       | 1.2421            | 0.1712              | 0.2053   |
| Number Holes        | 0.4667       | 1.5947            | 0.1963              | 0.0174   |
| Number Protrusions  | 0.3618       | 1.4359            | 0.1926              | 0.0603   |

*Bolding signifies features with p-value ≤ 0.05
| Tool           | Description/Objectives                                                                 | Data # of Shape Features | Applications          | Latest Version | Article | Package | Web Tool | Citation |
|---------------|----------------------------------------------------------------------------------------|--------------------------|-----------------------|----------------|---------|---------|----------|----------|
| EBImage       | This package provides general purpose functionality for image processing and analysis, | 4.36.0 (2021-12-19)      | ✓ ✓                   |                | ✓       |         | ✓        | [1]      |
| SAFARI        | This package provides functionality for image processing and shape analysis in the    | 0.1.0 (2021-02-25)       | ✓ ✓ ✓                 |                | ✓       |         | ✓        | [2]      |
| shapes        | This package offers routines for the statistical analysis of landmark shapes.          | 1.2.6 (2019-02-20)       | ✓                     |                | ✓       |         | ✓        | [3]      |
| wvtool        | This package intended to facilitate preprocessing and analyzing wood images.          | 1.0.0 (2016-11-08)       | ✓                     |                | ✓       |         | ✓        | [4]      |
Fig S1: An example of a whole-slide image from the National Lung Screening Trial (NLST) cohort processed by an Automated Tumor Recognition System Wang et al. (2018) and then converted into a binary format. The images are whole-slide pathology image (top), segmented three-class image (middle), and segmented two-class or binary image (bottom).
**Fig S2:** An example of the binary matrix (left) and its corresponding polygonal chain (right). The polygonal chain also shows the starting point and four sample points to demonstrated the contour’s clockwise direction.
\[ \mathbf{R} = \mathbf{R}' \times \mathbf{P} = \mathbf{P}' \times (\# \, \mathbf{r} = \mathbf{r}' \times \%) \]

---

**Fig S3:** Dependencies of shape features and representations. Colored boxes refer to the shape feature categories. Orange denotes geometric, green boundary, and grey topological.
**Fig S4:** Results from the regularized Cox proportional-hazards (CoxPH) model in **Downstream Analysis II: Predictive Performance.** The left figure shows the mean cross-validated errors, based on the Partial Likelihood Deviance, for each tuning parameter. The right figure shows the importance of each feature kept by the regularized Cox model, based on the magnitude of each coefficient.
REFERENCES

AGU, E. (2014). Digital Image Processing (CS/ECE 545) Lecture 8: Regions in Binary Images (Part 2) and Color (Part 1). Accessed: May 13, 2021.

BRADEN, B. (1986). The Surveyor’s Area Formula. The College Mathematics Journal 17 326–337.

GONZALEZ, R. C., WOODS, R. E. and EDDINS, S. L. (2020). Digital image processing using MATLAB, Third edition ed. Gatesmark Publishing, Knoxville.

KILDAY, J., PALMIERI, F. and FOX, M. D. (1993). Classifying mammographic lesions using computer-ized image analysis. IEEE Transactions on Medical Imaging 12 664–669.

POHLMAN, S., POWELL, K. A., OBUCHOWSKI, N. A., CHILCOTE, W. A. and GRUNDFEST-BRONIATOWSKI, S. (1996). Quantitative classification of breast tumors in digitized mammograms. Medical Physics 23 1337–1345.

WANG, S., CHEN, A., YANG, L., CAI, L., XIE, Y., FUJIMOTO, J., GAZDAR, A. and XIAO, G. (2018). Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. Scientific Reports 8 10393.

WIRTH, M. A. (2004). Shape Analysis & Measurement. Accessed: May 13, 2021.