Using Information About the Influence of Nutrient Diffusion On Tumor Evolution to Enhance Learning of Patient Trajectories

Hunter Bryan Boyce[1]*
Parag Mallick[1]
[1]Canary Center at Stanford for Cancer Early Detection
3155 Porter Drive
Palo Alto, California 94304
paragm@stanford.edu

Abstract

One of the critical challenges in cancer management is the prediction of which patients are likely to have aggressive disease that will require treatment. One emerging approach attempts to predict patient's likely trajectories from multiplex images of the expression of dozens of proteins within tumor biopsies. The majority of existing approaches take a pixel or cell-centric approach to prediction. The high dimensionality of these data representations alongside the restrictively small amounts of data available in most clinical studies lead to significant overfitting and generally poor overall performance. To address these challenges, we take advantage of our prior knowledge of the fundamental physical and evolutionary processes that drive tumor behavior. In particular, nutrient diffusion is a tremendously influential process that can lead some tumors to acquire extremely aggressive phenotypes. To incorporate this knowledge into an appropriate learning framework, we use a physics-motivated approach to focus attention to hypoxic regions of tumors. Focusing improves prediction performance and provides a foundation for future approaches to incorporate physics-based knowledge into learning approaches.

A major goal of cancer biology and clinical oncology is predicting the trajectory of a patient's disease as it can inform how to treat a particular patient. While we tend to think of cancer as a death sentence, the reality is that of the nearly 1.8M new cancer cases in 2020, average mortality rates are approximately 33% (Siegel, Miller, and Jemal 2020). However, Some cancers, such as cancers of the breast and prostate, have mortality rates below 20 percent. Unfortunately, the vast majority of learning approaches have struggled greatly with predicting which patients have aggressive disease (Kourou et al. 2015). Consequently we treat the entire population exposing many patients to the unnecessary toxicity and trauma of intervention. While treatment is required for patients with aggressive tumors, patients with more benign tumors may be best served by a watchful-waiting approach (Cooperberg et al. 2020).

*Special thanks to B. Anana for their helpful comments on this manuscript.
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Recently, there have been numerous examples of machine learning tools ability to identify objects in images and to classify images (e.g. cats and dogs) (Russakovsky et al. 2015). Typically, no prior, or mechanistic information is used in such classification problems. Likewise, there are no mechanistic functional forms governing the shapes in the images. However, through a combination of pooling and fully connected layers, the space of relationships is massively expanded to ultimately allow for the creation of extremely complex functions relating the input data to the terminal classes. Consequently, such methods typically require massive amounts of data to stably train. For some domains it is possible to gather millions of images of sufficient diversity to readily perform, with high accuracy, the two-way classification of images. For domains where data collection and labeling is complex or domains that are noisy, display high non-linearity, multistability, time evolution or multiscale behavior (e.g. pathological assessment of tumors), the majority of approaches struggle to extract the relevant features to appropriately learn the key drivers of system behavior (Zhu et al. 2020).

In high dimensional systems, like biological ones, there may be hundreds of thousands of moving parts and a lack of sufficient labeled training data. In addition, biology is controlled by processes like diffusion, which is dependent upon the temperature, viscosity and chemical composition of the environment and generates shapes that do not have the clear boundaries found with cats and dogs. Likewise the range of shapes that may exist, vary significantly. In addition to strict diffusion, there are interconnected issues of environmental heterogeneity, uptake, variation in density, variation in cell types, variation in vascular density and perfusion, etc. that challenge existing learning approaches. One possible solution to these challenges is to explicitly encode prior knowledge within the learning method. Cancer represents an ideal training ground for developing methods to best incorporate physics-based prior information and examine if doing so is able to enhance predictive power.

Diffusion as a Driver of Behavior in Cancer

Typical diffusion processes are modeled as a series of reaction-diffusion equations where diffusion is nearly in-

dependent of position. However, this is clearly not accurate in tumors. Tumors have non-uniform structure, variable compactness, and variable density (Dagogo-Jack and Shaw 2018). To take this structure into account, it is possible to assume that the viable cell density is proportional to the nutrient supply system and that the local oxygen consumption is proportional to cell density. Consequently, naturally occurring nutrient gradients resulting from diffusion limitations can create individual niches in which cells are typically exposed to either normoxic or hypoxic conditions.

Hypoxia, marked by a lack of oxygen in the tissue, is a phenotype of particular interest in cancer and leads to a more aggressive prognosis through hypermethylation (Thienpont et al. 2016), angiogenesis regulation (Pouyssgur, Dayan, and Mazure 2006), and remodeling of the extracellular matrix (Pouyssgur, Dayan, and Mazure 2006; Rankin and Giaccia 2016; Gilkes, Semenza, and Wirtz 2014). In order to measure hypoxia in the tissue, the levels of hypoxia responsive proteins, such as HIF1α or CA9, are quantified and used as a proxy for aggressivity. This methodology has proved useful in prognosis of cancer disease (Vaupel and Mayer 2007; Andersen et al. 2011; McDonald et al. 2012). However, quantification of protein alone does not accurately describe hypoxia in the tissue. Hypoxia is a result of unregulated growth leading to tissue being further from the blood vessels than the distance nutrients and oxygen can diffuse through the tissue (Frieboes et al. 2015). This makes hypoxia an inherently spatial process and it necessitates an understanding of the spatial context of cells in the tissue to understand how the expression of proteins in individual cells is related to the hypoxic process. While it isnt possible to readily measure the actual diffusion of oxygen through a tumor, it is possible to measure the expression of particular proteins, the locations and types of different cell types and accordingly the densities of cells. Some of those proteins (e.g. CA9) are extremely closely coupled to underlying diffusion processes.

We hypothesized that it may be possible to use these secondary measures to improve the learnability of higher-order predictors (e.g. patient outcome) and reduce the amount of data required for effective learning. Here we present an algorithm for identifying and focusing attention on regions of hypoxia in cancer tissue. Our method considers physical constraints and utilizes the spatial context and protein expression that is closely linked to the underlying diffusion of oxygen to quantify hypoxic regions. We apply this algorithm to a hyperplexed *in situ* data set of two-dimensional breast cancer histology samples (Jackson et al. 2020) and contrast our analysis with a pixel-based non-constrained analysis. We use the expression of CA9 as a proxy for hypoxia (figure 1a) as its expression is both associated with prognosis (Swinson et al. 2003) and spatially concordant with hypoxia in the tissue (Bao et al. 2012). We then demonstrate how using such an approach is able to improve learning over canonical approaches for multi-layer image data.

Method

Data modality

Measuring protein expression in the tissue (*in situ*) is typically done by immunostaining in which a reporter is conjugated to an antibody that is targeted to a protein of interest. We directly quantify the amount of reporter at specific locations as a proxy for the level of expression of a given protein at a given location. These reporters can be, but are not limited to, fluorophores (Gerdes et al. 2013), DNA oligonucliotides (Goltsev et al. 2018), or heavy metal ions (Angelo et al. 2014). Applied to cancer tissue samples, the result of such experiments give us a pixel-by-pixel annotation of protein expression across the sample. Cell borders are identified by segmentation and protein expression averaged over the cell regions to get whole cell expression of protein. This process gives us a data frame in which each row corresponds to an individual cell and the columns correspond to the sample the cell came from, the X and Y coordinates of that cell, and the expression of different proteins in that cell. For this analysis we are using a public dataset published earlier this year using imaging mass cytometry on a cohort of 289 breast cancer samples (Jackson et al. 2020). We use CA9 expression as a reporter associated with hypoxia (figure 1a).

Identifying hypoxia in the tissue

Hypoxia is a phenotype of the tissue, i.e. it is the direct result of physical laws governing diffusion of nutrients across space. Single pixels, or single cells alone cannot not describe this process. Therefore, to accurately measure hypoxia *in situ* and directly address hypotheses linking hypoxia and prognosis, we need to apply a transformation to the single cell spatial data. Specifically, we want to identify dense regions of tissue with high expression of CA9.

The hypoxia attention focusing algorithm has two main steps: a density-based clustering step and a segmentation step. The density-based clustering step is based on the density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al. 1996) and the segmentation step is based on alpha shapes.

Density-based clustering step

The DBSCAN algorithm has two parameters: `minPts` is specified as the minimum number of cells necessary to define a dense region and `epsilon` is specified as the radius around each cell. Input to the DBSCAN algorithm is the X and Y coordinates of the cell centers for cells that have CA9 expression higher than the 75th percentile (figure 1b).

The `minPts` parameter is determined by the sphere packing problem. In two-dimensions the number of non-overlapping circles that can surround a circle is 6 and we set `minPts` to this value. Because hypoxia is density dependent, this limits the size of possible identified regions through the algorithm.

The `epsilon` parameter is specified by the kth-nearest neighbor distance between the weighted centers of each cell in the sample. We set k to be the `minPts` value of 6. We divide each sample into a 5-by-5 grid and find the median density tile determine by total number of cells. This is done

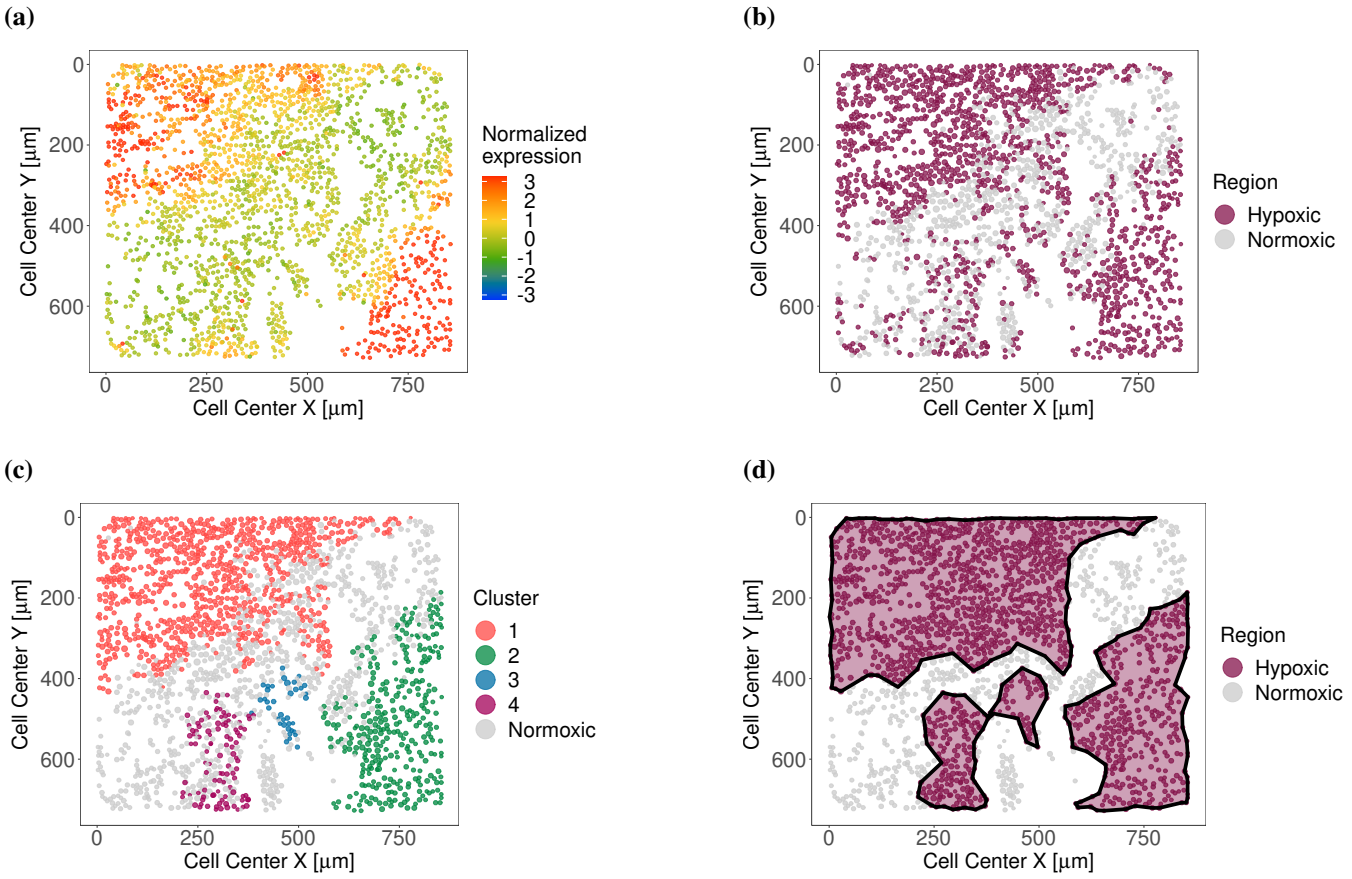

Figure 1: Visualization of algorithm steps. (a) Expression of CA9 protein in individual cells. (b) Individual cells that pass hypoxic threshold. (c) Clusters identified in DBSCAN algorithm. (d) Hypoxic regions identified by alpha shapes.

to limit the effect of extreme density differences within samples and to speed up computation. For the cells in the median density tile, the 6th-nearest neighbor distances calculated are sorted in increasing order and smoothed using locally weighted linear regression (LOESS) with a span of 5. We then approximate the curvature by finite differences and the distance at which the curvature is maximized is selected as `epsilon`. Cellular density can vary across the sample population because of cancer stage and because of variable breast density. Therefore, this procedure is done on a sample-by-sample basis.

To find regions of hypoxia, the DBSCAN algorithm is run on the subset of cells in each cancer sample that have expression of CA9 above the 75th percentile (considered to be highly likely to be experiencing hypoxia) with parameters `minPts` and `epsilon` as described above. The DBSCAN algorithm identifies core points (at least `minPts` within distance `epsilon`), border points (within distance `epsilon` of a core point [reachable]), and noise points (all points that are not reachable by any other point). We treat core points and border points, equally as part of a cluster. Noise points are discarded and not considered part of any cluster. Any clusters that are small (fewer than 25 cells) are also discarded as noise. The result of the density-based clustering step are shown in figure 1c.

Alpha shapes step

For each cluster defined by the density-based clustering step, we apply the Delaunay triangulation. The Delaunay triangulation is the set of all simplices (triangles in two-dimensions) such that no point in the point set is within the circumcircle of any simplex. The union of triangles formed by the Delaunay triangulation is the convex hull. However, the point set formed by the DBSCAN clusters are rarely convex and to account for this we must remove triangles from the triangulation. To prune triangles, we set a maximum threshold for the circumradius of each triangle to remain in the set and set the threshold to be `epsilon` from the density-based clustering step. The union of the triangles remaining after thresholding forms the alpha shape. Any cell that resides within the borders of an alpha shape, regardless of its CA9 expression is considered part of a hypoxic region (1d).

Results

Examining the cells that are considered hypoxic with no spatial prior information (figure 1b) we observe that non-

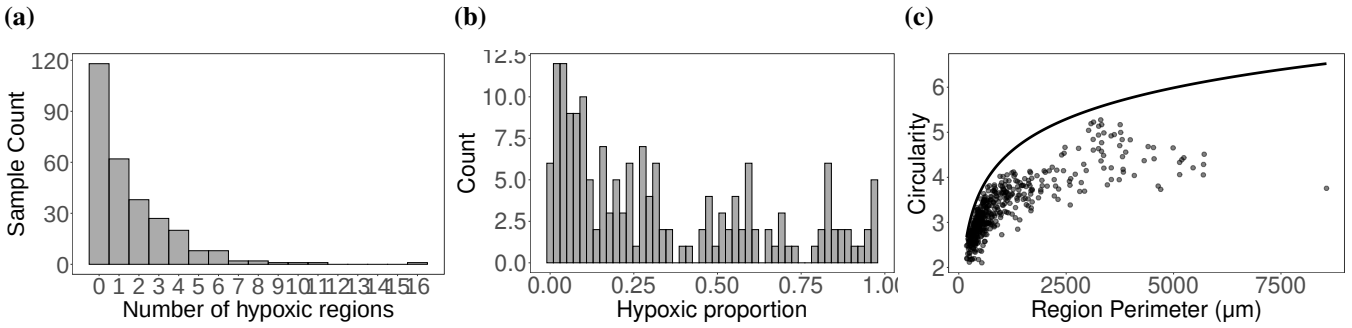

Figure 2: Physical properties of hypoxic regions in the tissue. (a) Histogram of the number of hypoxic regions identified across samples. (b) Histogram of the proportion of hypoxia identified across samples. (c) Circularity of the hypoxic regions identified. Circularity is defined as the log ratio of area to perimeter for each region. Curved solid line corresponds to the circularity of a circle at a given perimeter.

Before → After	Count	Percent
Normoxic → Normoxic	515496	89.5%
Normoxic → Hypoxic	60378	11.5%
Hypoxic → Normoxic	39362	20.5%
Hypoxic → Hypoxic	152597	79.5%

Table 1: Counts of cell type transitions before and after application of algorithm

Model	coef	S.E.	χ^2	p-value
Percent positive pixels	-1.8762	0.9131	4.22	0.039
Percent positivity	-0.9097	0.4338	4.4	0.036
Percent area	-0.9116	0.3788	5.79	0.0161
Circularity	-0.1471	0.0646	5.18	0.0229

Table 2: Performance of accelerated failure time models on patient survival

constrained hypoxia identification results in regions that are quite sparse and not representative of regions of hypoxia in the tissue. Contrasting that to hypoxic cell identification with spatial constraints (figure 1d), we obtain only appropriately dense regions that more accurately represent the diffusion driven hypoxic phenotype. Similarly, we observe that normoxic cells with no constraints fall within hypoxic regions when considering the spatial context. The cells that are lost to sparsity are 20.5% of the hypoxic cells and the cells that fall within a hypoxic region are 11.5% of the normoxic cells (table 1).

Almost every sample contains hypoxic cells with no spatial constraints (288 of 289), but after consideration of the spatial context only 171 of the samples have hypoxic regions (figure 2a). Most of samples contain between 1 and 4 regions and a small number of samples that have a high amount of spatial heterogeneity with 5 or more regions. Of those samples with spatially constrained hypoxic regions, most of them take up less than half of the sample area (figure 2b).

We wanted to quantify how irregular the regions identified with our algorithm were. To achieve this we calculated the log of the ratio between the area of each spatially constrained hypoxic region to its perimeter (figure 2c). We observe that regions with smaller perimeters are more circular (closer to the curved line). As perimeters get larger, there are a wider range of conformations. However, the largest perimeters have very irregular shapes, reflecting the heterogeneous tumor microenvironment that influences the diffusion of nutrients.

To test the clinical utility of encoding the physical constraints of hypoxia in the tissue, we conducted a survival analysis with an accelerated failure time model. The response is survival in number of months post-diagnosis and is right censored. We fit four univariate models: first we used percent of positive pixels from the images; second we used percent positivity (proportion of hypoxic cells with no spatial constraints); third we used percent area (proportion of sample area affected by hypoxia), and fourth we used median circularity (2c). We use the exponential distribution as determined by fitting a null model to the survival data. While, all models show significance (as determined by an α threshold of 0.05) the standard error of the percent area model and the circularity model (both consider physical constraints) is smaller than the percent positivity models (no constraints) (table 2) demonstrating that diffusion inspired prior information can improve prediction performance.

Conclusion

Traditional learning techniques on image data, or on multilayer image data take a pixel-centric approach and suffer from huge challenges in overfitting and predictive performance. In our study, we recognize the importance of higher-order spatial features, as created by the physical process of nutrient diffusion throughout a tumor. Additionally, focusing on individual cell as the atomic unit (as opposed to individual pixels) and by incorporating prior knowledge of this physical phenomena through an attention focusing approach we demonstrate improvements in prediction performance wtihin a relatively small cohort of labeled train-

ing data and have more biologically interpretable results. Given the emerging availability of rich datasets and algorithms, such as ours, that appropriately consider the spatial context governing the progression of cancer, we foresee a greater understanding of cancer that goes beyond that of the individual cell. In addition, we note that physics-based attention focusing approaches have implications to other modeling techniques, such as defining focus regions for CNN-based approaches, or defining the shapes of the convolution functions themselves. Ultimately, we anticipate that physics-based approaches to incorporate prior information into learning methods will allow us to leverage decades of theoretical work as priors to improve learning.

References

Andersen, S.; Eilertsen, M.; Donnem, T.; Al-Shibli, K.; Al-Saad, S.; Busund, L.-T. T.; and Bremnes, R. M. 2011. Diverging prognostic impacts of hypoxic markers according to NSCLC histology. *Lung cancer (Amsterdam, Netherlands)* 72(3):294–302.

Angelo, M.; Bendall, S. C.; Finck, R.; Hale, M. B.; Hitzman, C.; Borowsky, A. D.; Levenson, R. M.; Lowe, J. B.; Liu, S. D.; Zhao, S.; Natkunam, Y.; and Nolan, G. P. 2014. Multiplexed ion beam imaging of human breast tumors. *Nature medicine* 20(4):436–42.

Bao, B.; Groves, K.; Zhang, J.; Handy, E.; Kennedy, P.; Cuneo, G.; Supuran, C. T.; Yared, W.; Rajopadhye, M.; and Peterson, J. D. 2012. In vivo imaging and quantification of carbonic anhydrase IX expression as an endogenous biomarker of tumor hypoxia. *PloS one* 7(11):e50860.

Cooperberg, M. R.; Zheng, Y.; Faino, A. V.; Newcomb, L. F.; Zhu, K.; Cowan, J. E.; Brooks, J. D.; Dash, A.; Gleave, M. E.; Martin, F.; Morgan, T. M.; Nelson, P. S.; Thompson, I. M.; Wagner, A. A.; Carroll, P. R.; and Lin, D. W. 2020. Tailoring Intensity of Active Surveillance for Low-Risk Prostate Cancer Based on Individualized Prediction of Risk Stability. *JAMA Oncology* 6(10):e203187.

Dagogo-Jack, I., and Shaw, A. T. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* 15(2):81–94.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. 96(34):226–231.

Frieboes, H. B.; Curtis, L. T.; Wu, M.; Kani, K.; and Mallick, P. 2015. Simulation of the Protein-Shedding Kinetics of a Fully Vascularized Tumor. *Cancer informatics* 14:163–75.

Gerdes, M. J.; Sevinsky, C. J.; Sood, A.; Adak, S.; Bello, M. O.; Bordwell, A.; Can, A.; Corwin, A.; Dinn, S.; Filkins, R. J.; Hollman, D.; Kamath, V.; Kaanumalle, S.; Kenny, K.; Larsen, M.; Lazare, M.; Li, Q.; Lowes, C.; McCulloch, C. C.; McDonough, E.; Montalto, M. C.; Pang, Z.; Rittscher, J.; Santamaria-Pang, A.; Sarachan, B. D.; Seel, M. L.; Seppo, A.; Shaikh, K.; Sui, Y.; Zhang, J.; and Ginty, F. 2013. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences* 110(29):11982–11987.

Gilkes, D. M.; Semenza, G. L.; and Wirtz, D. 2014. Hypoxia and the extracellular matrix: drivers of tumour metastasis. *Nature reviews. Cancer* 14(6):430–9.

Goltsev, Y.; Samusik, N.; Kennedy-Darling, J.; Bhate, S.; Hale, M.; Vazquez, G.; Black, S.; and Nolan, G. P. 2018. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 174(4):968–981.e15.

Jackson, H. W.; Fischer, J. R.; Zanotelli, V. R.; Ali, R. H.; Mechera, R.; Soysal, S. D.; Moch, H.; Muenst, S.; Varga, Z.; Weber, W. P.; and Bodenmiller, B. 2020. The single-cell pathology landscape of breast cancer. *Nature* 1–6.

Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; and Fotiadis, D. I. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13:8–17.

McDonald, P. C.; Winum, J.-Y. Y.; Supuran, C. T.; and Dedhar, S. 2012. Recent developments in targeting carbonic anhydrase IX for cancer therapeutics. *Oncotarget* 3(1):84–97.

Pouyssgur, J.; Dayan, F.; and Mazure, N. M. 2006. Hypoxia signalling in cancer and approaches to enforce tumour regression. *Nature* 441(7092):437–43.

Rankin, E. B., and Giaccia, A. J. 2016. Hypoxic control of metastasis. *Science (New York, N.Y.)* 352(6282):175–80.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115(3):211–252.

Siegel, R. L.; Miller, K. D.; and Jemal, A. 2020. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* 70(1):7–30.

Swinson, D. E.; Jones, J. L.; Richardson, D.; Wykoff, C.; Turley, H.; Pastorek, J.; Taub, N.; Harris, A. L.; and O'Byrne, K. J. 2003. Carbonic anhydrase IX expression, a novel surrogate marker of tumor hypoxia, is associated with a poor prognosis in non-small-cell lung cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 21(3):473–82.

Thienpont, B.; Steinbacher, J.; Zhao, H.; D'Anna, F.; Kuchnio, A.; Ploumakis, A.; Ghesquire, B.; Dyck, L. V.; Boeckx, B.; Schoonjans, L.; Hermans, E.; Amant, F.; Kristensen, V. N.; Koh, K. P.; Mazzone, M.; Coleman, M.; Carell, T.; Carmeliet, P.; and Lambrechts, D. 2016. Tumour hypoxia causes DNA hypermethylation by reducing TET activity. *Nature* 537(7618):63–68.

Vaupel, P., and Mayer, A. 2007. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer metastasis reviews* 26(2):225–39.

Zhu, W.; Xie, L.; Han, J.; and Guo, X. 2020. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers* 12(3):603.