Introduction

The diagnosis of cancers in modern medicine typically involves a biopsy of the tissue in question followed by visual examination by a licensed pathologist (McKenney, 2017). The nature of this process gives way to improvements in efficiency through automation, particularly through the use of computer vision algorithms. One recent example of this ideology put into practice is the development of a Convolutional Neural Network (CNN) that can produce dermatologist-level accuracy in the classification of skin cancer images (Esteva et al., 2017). Developments of this nature can bring improvements to medicine and patient care through enhancements in efficiency and reduction of costs.

While there has been much progress recently in computer-aided diagnosis in areas including radiology and dermatology, there have been fewer efforts in the development of algorithms that can effectively classify histopathology slides. Technologically-focused works in the area of histology have focused primarily on a single cancer type (Cireşan et al., 2013; Xu et al., 2016); it is likely that work in this arena has been limited by the availability of high quality data sets. The project proposed here overcomes this barrier by using data from the Cancer Digital Slide Archive – a high quality collection of curated histopathology image slides representing 31 disease classes.

Data

Data will be obtained primarily from the Cancer Digital Slide Archive (CDSA) (Gutman et al., 2013). This archive houses histopathology slides for biopsied cancerous tissue, originating from 31 cancer types (a full list can be found in **Appendix A**). Collected data will be comprised of mainly images, and would include some basic accompanying metadata (i.e. disease class). The

data is stratified by disease class, patient, and slide type. Slide types include Diagnostic (DX), and Permanent (TS).

Images for this project will exist in two types, termed here as "Slide Level" and "Tile Level". Images on CDSA are high-resolution mosaics, which makes the acquisition of multi-resolution tiles possible. For this project, tiles at a level of 15x was chosen as this was deemed to appropriately show detail at the cellular level. An example of these image types is outlined here:

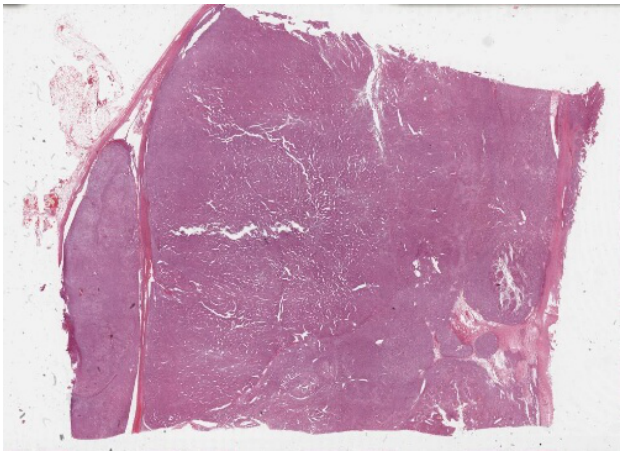

Fig 1a. "Slide Level": Diagnostic section of Adrenocortical carcinoma tissue

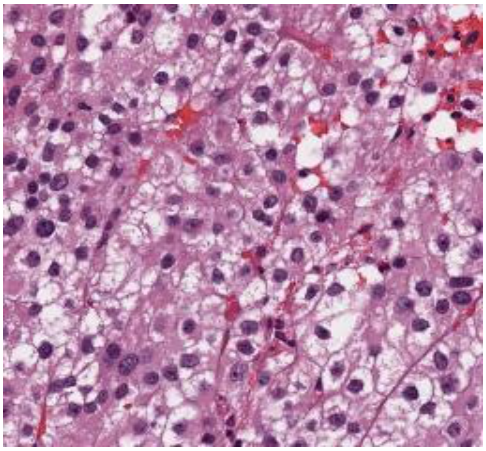

Fig 1b. "Tile Level": 15x zoom tile taken from **Fig 1a.**

Goals

The ultimate goal of this project is the development of a computer vision algorithm that can correctly classify a given histology image into one of 31 given classes. Focuses of this project include:

Using both Slide Level and Tile Level images:

1) Comparison of *de novo* neural network architectures, naïve Google Inception v3 architecture, and pre-trained Google Inception v3; the latter will involve reimplementation of the output layer of the network

Using only Slide Level images:

2) Accurate Classification of DX and TS sections for a given tumor type
3) Differentiation between DX and TS sections for tumor types other than the ones used for training (generalizability)

Using both Slide Level and Tile Level images:

4) Differentiation between tumor types that are known to look similar to humans. For example:
    a. Lung Adenocarcinoma vs. Lung Squamous Cell Carcinoma
    b. Kidney renal clear cell carcinoma vs . Kidney renal papillary cell carcinoma
    c. Glioblastoma multiforme vs. Brain Lower Grade Glioma

Data Transformation

Tile Level data: During the scraping process, tiles will be requested in 256x256 dimensions. These images may be scaled down as experimentation progresses. Color may need to be normalized.

Slide Level data: Slide Level images come in very different sizes based on the size of the tissue biopsied. As such, these images will need to be transformed into uniform dimensions, either through stretching or through padding. Color may need to be normalized.

Technologies

Python will be used for most tasks, along with the Keras framework for machine learning functionality.

References

- Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013, September). Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (pp. 411-418). Springer, Berlin, Heidelberg.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118.
- Gutman, D. A., Cobb, J., Somanna, D., Park, Y., Wang, F., Kurc, T., ... & Kong, J. (2013). Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *Journal of the American Medical Informatics Association*, *20*(6), 1091-1098.
- McKenney, J. K. (2017). The present and future of prostate cancer histopathology. *Current Opinion in Urology*, *27*(5), 464-468.
- Xu, J., Luo, X., Wang, G., Gilmore, H., & Madabhushi, A. (2016). A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, *191*, 214-223.

Appendix A – Disease States

Abbr.	Name
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma