Medical Image Processing Project

Cancer Clinical Decision Support Enabled by Medical Image Processing of Histopathological Images from Cancer Patients

Introduction

The goal of this project is to assist students in learning the complete translational biomedical image processing pipeline for clinical decision support. Specifically, the students will apply image processing and data mining techniques to cancer histopathological images, and develop an objective and reproducible decision support for diagnosis and prognosis of cancers. The cancer images from The Cancer Genome Atlas (TCGA) will be provided to student to accomplish this goal. TCGA is a joint project by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The goal of the TCGA project is to accelerate the understanding of cancer in order to more effectively diagnose, cure, and prevent cancer. TCGA provides high-quality genomic, proteomic, and imaging data and students will need to sign Data Access Requirement Form before giving access to these data. The TCGA cancer imaging data include whole slide tissue biopsy samples and MRI data for multiple types of cancer, where more than 1000 whole-slide images (~20,000×40,000-pixel in size) are provided for each cancer. Students will be provided sub-section images (512×512 pixels) that have been preprocessed after “quality control” procedure from few WSIs. The student teams will work on image segmentation, feature extraction, data mining, etc, which form the complete problem solving workflow in building a clinical decision support system.

Structure

This project is divided into three modules:
(1) Image preprocessing.
(2) Image feature extraction
(3) Image classification

The modules are sequential blocks of a translational image processing pipeline for clinical decision making, and as such, methods developed in a module will be used by subsequent modules.

Project Module 1: Image Preprocessing

The goal of this module is to perform image preprocessing on histopathological images. Image preprocessing includes quality control (ink removal, tissue folding removal, etc.), data augmentation and color/stain normalization.

Data

300 digital microscopic images of hematoxylin and eosin (H&E) stained tissue sections of kidney clear cell carcinoma consisting of 100 tumor, 100 necrosis, and 100 stroma sections will be provided. H&E staining enhances three colors: blue-purple, white, and pink. These colors correspond to specific cellular structures.

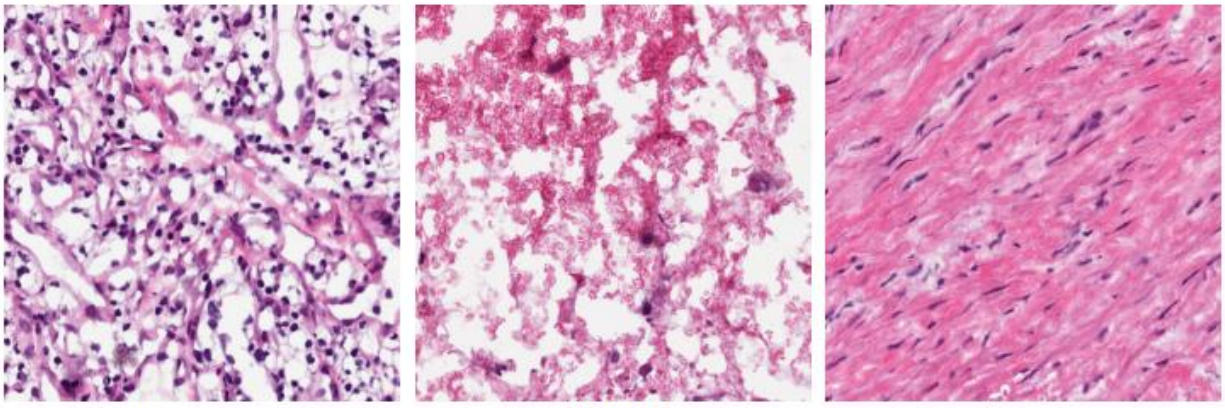

Fig. 2. Example images in the kidney data set

Most researchers segment nuclear structures based on color properties. Thus, each image consists of three parts:

1) Basophilic structures containing nucleic acids—ribosome and nuclei—tend to stain blue-purple;
2) Eosinophilic intra- and extracellular proteins in cytoplasmic regions tend to stain bright pink;
3) Empty spaces—the lumen of glands—do not stain and tend to be white.

In traditional machine learning pipeline [1,2,3], researcher would do image segmentation to identify nuclei location and density. Nuclei features will later used for classification purpose. In this project, however, we skip image segmentation, mainly due to the lack of pixel-level annotation for evaluation.

In this module, you are expected to do color normalization and

Requirements:

1. Implement color normalization using “Reinhard’s Method” [4].
2. Manually code data augmentation, including random crop (224*224 pixels), flipping and rotation.

Reference:

[1] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, et al., "Automatic batch-invariant color segmentation of histological cancer images," Conf Proc IEEE Int Symp Biomed Imaging, ISBI, pp. 657-660, 2011.
[2] S. Di Cataldo, E. Ficarra, A. Acquaviva, E. Macii, “Achieving the way for automated segmentation of nuclei in cancer tissue images through morphology-based approach: A quantitative evaluation,” Computerized Medical Imaging and Graphics 34 (2010) 453–461
[3] H. Zhang, E. Fritts Jason, and A. Goldman Sally, "Image segmentation evaluation: A survey of unsupervised methods," Computer Vision and Image Understanding, vol. 110, pp. 260-280, 2008.
[4] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, et al., "Colour Normalisation in Digital Histopathology Images," Proc. Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop), pp. 100-111, 2009.

Project Module 2: Feature Extraction and Selection

The goal of this module is to extract and explore image features in histopathological images. You will develop and compare different feature extraction and feature reduction methods.

Requirements:

1. Conduct literature survey on histopathological image analysis. Find **at least one paper** about feature extraction on histopathological images **by your own**. Feel free to read whichever paper in [5-10]. Summarize **at least 3 types of features** reported for histopathological images.
2. Implement algorithms to extract three categories of image features (more than 100 features in total) as reported in [5-10]:
    a. Color
    b. Texture such as wavelet, GLCM, and fractal
    c. Morphological features
3. Implement/Apply data dimensionality reduction algorithms, such as PCA. Discuss the loss of information when samples are represented by reduced data dimension in the transformed domain.
4. If you can implement a simple autoencoder or convolutional neural network to extract features, you will get bonus points.

Reference:

[5] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review," *IEEE Rev Biomed Eng*, vol. 2, pp. 147-171, 2009.
[6] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image feature mining reveals emergent diagnostic properties for renal cancer," *Conf Proc IEEE Bioinform Biomed*, *BIBM*, 2011.
[7] L. Boucheron, "Object-and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer," PhD thesis, University of California, Santa Barbara, 2008.
[8] J. Barker, A. Hoogi, A. Depeursinge, D. L. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Medical Image Analysis*, vol. 30, pp. 60-71, 2016.

[9] P. Huang and C. Lee, "Automatic Classification for Pathological Prostate Images Based on Fractal Analysis," in *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1037-1050, July 2009.
[10] H. Wang, A.C. Roa et al, “Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features” . *Journal of Medical Imaging*, 1(3), 034003 (2014).

Project Module 3: Prediction Modeling

The goal of this module is to develop and validate computer-based prediction models. This is the foundation for clinical diagnosis decision support system development based on cancer biopsy images. You will build prediction models using different classifiers on training dataset to train the model and validate the predictive models using test dataset. You shall apply cross-validation and utilize different performance metrics.

You will develop a multiclass classification model for this dataset.

Deliverables

For completion of this module, you will perform the following tasks:

1. Conduct literature survey on computer-based classifiers for image analysis **(at least four peer-reviewed papers)** to identify both **linear classifiers and non-linear classifiers**. Then summarize **at least four typical classifiers for decision making**. Examples include support vector machine (SVM), k-nearest neighbor (KNN), and random forest to show your understanding of the strength or limitation of each method.
2. Develop prediction models using **three classifiers** out of the classifiers you have surveyed and critiqued.
3. Develop **two cross-validation schemes** (or internal validation) by training the classifiers using the training dataset
    a. N-iterations of m-fold cross-validation
    b. Leave-one-out cross-validation or bootstrapping for classifier optimization
4. Develop **four performance metrics to evaluate** your predictive model performance when applying to test dataset
    a. Area Under the Curve (AUC)
    b. Matthews Correlation Coefficient (MCC)
    c. F1-score
    d. Accuracy