Kevin Brockman

Project Proposal

Introduction to Research Computing

Introduction

Automated classification of histology sections for cancer diagnosis has seen continuous improvement in recent years using machine learning methods. Traditional diagnosis is based largely on experience rather than strict rules meaning training into the role is time consuming. Machine learning is an obvious fit for this such an application as it can distill this training into a widely usable model to allow fast, accurate results for patients.

In a pilot study it was shown that a combination of raw imaging data (as opposed to more processed data with spatial profiling) and randomly sampled chemical signatures were effective at differentiating benign and malignant samples in breast cancer. This data was trialed using autoencoder, decision tree, and convolution neural network machine learning models with the autoencoder found to be the most effective and efficient (Winkelmaier, 2018). I would like to progress with work further to create an effective resource for detection of malignant cells.

Statement of Problem

The original study was limited in scope and must be expanded upon to further improve and verify the efficacy of these methods in the diagnosis of breast cancer. Of specific note are the following:

1. In the original study modelling was performed using a combination of an imaged histology section and randomly sampled spectral data to identify levels of certain chemical components
    a. A more robust model could be developed by aligning a full set of spectrographic data with the corresponding locations on the image
2. The original study did not investigate the efficacy of imaging and spectral data separately
3. The original study may have failed to achieve the full potential of convolutional neural networks by working from scratch rather than beginning training with a partially developed general image classification model
4. The original study worked off a very limited data set (22 samples)

Objectives

Practical Goals:

1. Develop basic models to differentiate benign and malignant cells in homogenous samples (containing only malignant or only benign in a given sample)
2. To develop advanced models to differentiate benign and malignant cells in mixed (heterogenous) samples (potentially containing both types of cell in an individual sample)
3. Optimize the above models and provide a robust analysis of their strengths and weaknesses
4. To leave the above work when it is completed or my time with it is done with sufficient documentation and intuitive design for future work to be picked up by another

Personal Goals:

1. Improve knowledge and skill with machine learning methods
2. Improve technical documentation skills
3. Improve knowledge of research methodology
4. Gain knowledge in cancer biology.

Plan of Action

Toolset:

Python will be the primary language used in this project for the following reasons

1. Python has a robust supply of machine learning libraries available for free use
2. Python is an accessible language that is commonly used in scientific pursuits, in the absence of someone dedicated to computation this will be relatively easy for the researcher of the lab to work with in the future
3. Potential gains in time from working on a purely compiled language have been determined to not make up for the advantages of python, the software being written will primarily be calling libraries likely written in compiled languages and not doing novel processing itself

Project Steps:

1. Individual Models
    a. Trial models using image data with different machine learning techniques (e.g. Encoder, convolutional neural network) and optimization within individual models
        i. Initial training with homogeneous data followed by further testing and optimization with heterogenous data
    b. Trial models using spectral data with different machine learning techniques (e.g. Encoder, convolutional neural network) and optimization within individual models
        i. Initial training with homogeneous data followed by further testing and optimization with heterogenous data
2. Composite Model
    a. A method will be developed to combine and align image data with full spectrographic data
    b. A model will be trained on the above data as in step one
3. Analysis
    a. The optimized models will be compared on three primary criteria
        i. Accuracy (and other statistical measures) of models to compare their effectiveness and the confidence in that effectiveness
        ii. Time and cost needed to acquire the input data for a given model (imaging vs spectrographic vs composite model vs combination of individual model results)
        iii. Test time for individual samples (if significant)
4. Potential Further Work
    a. Testing efficacy of methods on other cancers
        i. Possibly simply the method for developing models, or possibly as a base model that can be specialized into individual cancers similar to pretrained image classification models
    b. Develop toolkit for the training and application of models

Conclusions

Machine learning based pathology is not a novel concept but this particular approach using spectral data shows potential to improve the field. I will develop multiple models using the data of interest in order to identify their merits. It is my hope that this work will ultimately lead cheaper more effective diagnosis of cancers.

References

1. Winkelmaier, Garrett, et al. "Quantum Cascade Laser Infrared Microscopy Differentiates Malignant Phenotypes in Breast Histology Sections." *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, https://doi.org/10.1109/isbi.2018.8363700.