Kevin Brockman

Project Design

Summary

Image and spectrographic data of histology sections will be used to train a collection of machine learning models which will then be compared in an attempt to identify the ideal model for classifying samples as benign or malignant.

File Management

All software generated as part of this project will be made available to the lab via a git repository including instructions for the setup and use of the software in a readme file. The project will be stored in a primary directory with subdirectories including but not limited to: Training & test data, Samples for analysis, generated models, and software files. This primary directory will also include an executable file to analyze sample data using generated models. All generated software files will be commented to explain their internal functional blocks.

Toolset

The software will be developed in python. The project will include but will not be limited to the following libraries: numpy, sklearn, pandas, and tensorflow. Numpy and pandas provide general mathematical and data management functionality used with the other libraries. Sklearn and tensorflow provide the functionality for generating and testing the machine learning models.

Data Input & Preparation

Data will be received in its raw form from imaging. Each sample may contain visual image data, spectrographic data, or both that must be properly combined by the software. The fully combined data will be in the form of a 2D image where each pixel includes the original image coloration, and the 2D spectrographic information for that location. Other partially combined forms may be used as well for different models. If necessary for the development of certain models other data processing may be performed, including but not limited to smoothing, padding, and normalization.

Model Generation

Two categories of model will be the focus of analysis:

1. Auto-Encoder with softmax classifier, which was identified as the most promising candidate in a pilot study
2. Convolutional Neural Network (CNN), which is ideal for image classification

Input data will be provided to these models in 3 primary forms:

1. Image data
2. Spectrographic data
3. Combined image and spectrographic data

Depending on initial results for models, poorly performing models will be removed from consideration while well performing models will be further varied and tuned until an ideal state is reached based on evaluation criteria. So long as a model performs well in at least one criterion it may be maintained for its potential utility. Each model, its parameters, and its evaluation will be documented in a spreadsheet for future reference. The details of the parameters being manipulated will depend on the model in question but may include hidden layer size and number for CNN and degree & method of regularization for softmax classifier.

Initial development will be performed using simple homogeneous sample data (all benign or all malignant), after these have been confirmed to be effective more complex heterogenous samples with a mix of malignant and benign cells will be used.

Where possible this branching testing will be automated to work through generations of models selecting for the best performers, though care will need to be taken to prevent overfitting to the training data. This will involve generating some number of initial model variants from a base, generating and testing models, and dropping a certain number of poorer performers, then beginning again with variants of the high performers. All tested models should be documented for future reference and to prevent redundant work.

Model Analysis

Models will be evaluated based on statistical metrics for results, computer performance metrics, and the time & cost needed for the requisite imaging data (visual imaging, spectrographic data, or both). Including but not limited to the following.

1. Statistics
    a. P-value (Determined via two-tailed t-test)

b. Accuracy (% correct classification of training data)

c. Precision (% correct positive predictions of training data)

d. Recall (True positive rate)

e. F1 Score (Composite of precision and recall)

i. This will be used as the primary scoring method when incrementally improving models

2. Performance

a. Training time

i. For this application this is unlikely to be a significant factor and may be dropped from consideration if this is shown to be the case

b. Test time

i. For this application this is unlikely to be a significant factor and may be dropped from consideration if this is shown to be the case

3. Imaging

a. Expected time from sample collection to receiving images for processing

i. This factor will be based purely on which data are used in the generation of the model resulting in 3 possible categories (image, spectrographic, or both)

b. Expected cost for having image data gathered

i. This factor will be based purely on which data are used in the generation of the model resulting in 3 possible categories (image, spectrographic, or both)

Deliverables

When complete the best performing models of each type (classification methods and input data types) should be documented with an analysis of the strengths and weaknesses of each and any insights that may have been gleaned during development. This is in addition to the files described in the ‘File Management’ section.