DS 340W Project Proposal

Katherine Graham and Taylor Bostick

I. Purpose of the Project

THE purpose of this project is to create a data science solution to a real-world problem. This assignment is designed to provide hands-on training in data science problems and solutions. After working through and completing this project, the team, as well as the class as a whole, will have a deeper understanding of how problem-solving in the data science world takes place.

For our project we are focusing on how data science can be used to facilitate the detection of breast cancer. Making sure that we come up with a solution that is accurate is an important part of the project. Our team wants to use data science techniques to effectively detect breast cancer.

A. Who Will This Benefit?

Breast cancer is most commonly detected in women, and is one of the highest leading causes of death for women. Accurate detection means that treatment can begin sooner, which means it will have a higher success rate. As stated previously, we are focusing on using data science to detect breast cancer, which will benefit women by getting them into treatment, and hopefully, remission faster.

Additionally, this is an issue that women all over the world encounter. Our project could have the ability to benefit both women in the United States as well as across the world.

II. Technical Challenges

Aside from the challenge of developing an optimized model, we foresee two potential areas of technical concern.

The first involves finding a dataset with sufficient clarity on both the features within the dataset and the data source itself. Though breast cancer data is abundant, many of the initial datasets we reviewed contained little to no information about the fields captured. Some others included links to source pages that simply no longer existed.

The second area of technical concern involves settling on the right model for our data. The dataset we intend to work with includes digital pathology images of breast cancer screenings. Though there are a number of applicable models, the two candidates we are considering include decision trees (DTs) and convolutional neural networks (CNNs). DTs have the advantage of being more easily interpretable over any type of neural network, and they can handle large amounts of data.

On the other side, CNNs are ideal for training on complex images. Though a class of neural networks, CNNs fall under the umbrella of deep learning. They are advanced models best suited for image analysis tasks such as that outlined in our own project objective. However, due to our own lack of experience working with DL models, developing this model may prove to be a third, critical challenge in itself.

III. Data Science Solution Framework

Our ultimate goal with this project is to produce a model optimized to detect breast cancer in images. Despite our lack of experience working with DL models, a convolutional network will surely produce the best results. Due to this, we believe that our solution framework should involve a comparative analysis of model results for the ensemble decision tree method, random forests (RFs), and CNNs. Since we are still unsettled on a dataset, below is an outline of our foreseen solution framework:

Project Solution Framework

- Pre-processing Step
  - Assess data (e.g., outliers, missing values)
  - Data treatment (e.g., missing value treatment, ”noisy” value treatment)
  - Data transformation (e.g., normalization, feature selection)
- Feature Extraction Step
  - Extract significant features (e.g, linear discriminant analysis, PCA)
- Model Fitting / Classifier Step
  - Feed extracted features through base models
    - RFs: By means of cross validation, optimized tuning parameter returns optimized model that minimizes test prediction error
    - CNNs: Three main layers will need to be created: convolutional, pooling, and fully connected
      - Avoid overfitting through data augmentation
  - Model Evaluation Step
    - RF: Metrics includes classification report, confusion matrix, and accuracy scores
    - CNN: Metrics can be retrieved by means of automatic or manual verification datasets

IV. Implementation Plan

For Week 6 we hope to be finished with formulating the problem, collecting data, and cleaning/formatting the data. As a team we believe that we can have these steps completed for the Week 6 check in. Additionally, we hope to have started developing a solution, but realize that this will be time consuming and will not be done before Week 6, but we will have started this step

As for Week 9, the team expects to be finished developing our solution and close to finished with conducting experiments. After completing the experiments we will then be analyzing the results and beginning to formulate our findings into a presentation and final report.