Big Data Technology

CSCI 493.76: Project 3 – Breast Cancer Detection

Spring 2024

Grade = minimum(total score, 100).

Introduction

Breast cancer stands as the leading cancer diagnosis among women globally, comprising a quarter of all cancer instances, with over 2.1 million individuals affected in 2015 alone. It initiates when cells within the breast experience uncontrolled growth, typically resulting in the formation of detectable tumors visible through X-ray (mammograms) imaging or palpable as lumps in the breast region.

The primary obstacle in detecting breast cancer lies in effectively categorizing tumors as either malignant (cancerous) or benign (non-cancerous). You will undertake the analysis of tumor classification using machine learning techniques utilizing the Breast Cancer Wisconsin (Diagnostic) Dataset.

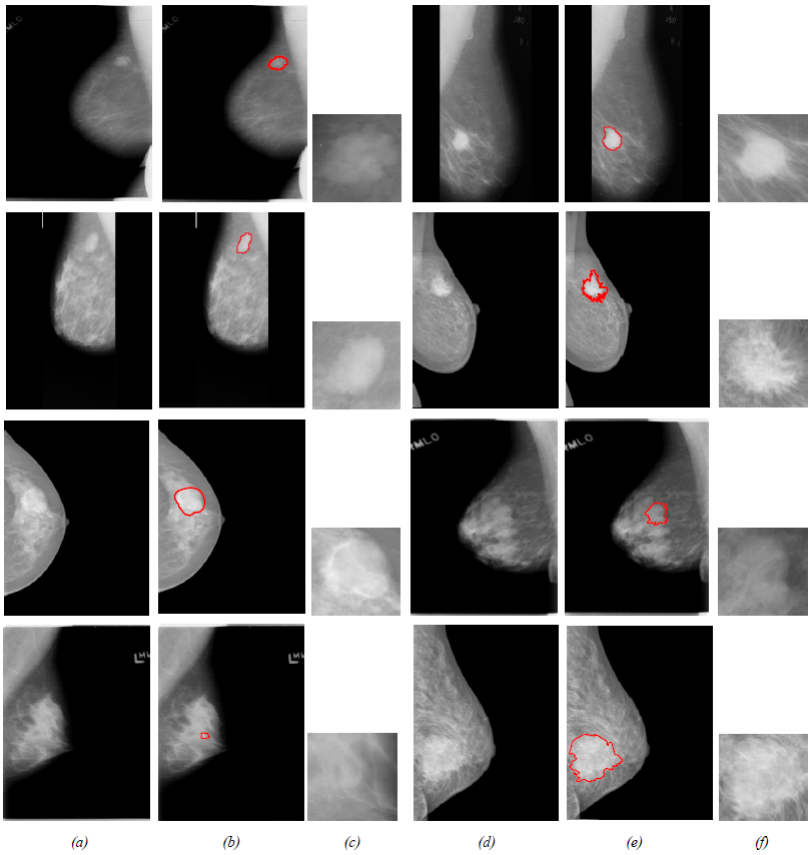

Figure 1: Detection of Breast Cancer Masses from Different Breast Density Categories; from top to bottom: Fatty, Glandular, Dense and Extremely Dense; (a, d) Original Images (b, e) Ground Truth Images (c, f) Extracted Patches by the proposed algorithm.

Coding Section (70 points)

1. (Readin file) Goto the following website to download the Breast Cancer Dataset. Read the CSV file into panda data frame. You can choose any data structure type to store the values (pandas Or Mat-Lab). 32 features (or columns).

2. (data cleaning / Preparation) Remove any row which contain empty cell(s) "Bad Data." Split the dataset into training set (80%) and testing (20%).

3. (Modeling and Evaluations) Train the dataset on the Decision Tree Classifier using the training set (Track the training time). Draw the decision Tree. Evaluate your trained model using the testing data. How well does your model perform? Use performance metrics, like accuracy, sensitivity and specificity (recall). Visualize the confusion Matrix.

4. (Modeling and Evaluations) Train the dataset on the Support Vector Machine (RBF) Classifier (Track the training time). Evaluate your trained model using the testing data. How well does your model perform? Use performance metrics, like accuracy, sensitivity and specificity (recall). Visualize the confusion Matrix.

5. (Teams with 3 people) Train the dataset on the Support Vector Machine (Polynomial) Classifier (Track the training time). Evaluate your trained model using the testing data. How well does your model perform? Use performance metrics, like accuracy, sensitivity and specificity. Visualize the confusion Matrix.

6. Find the feature importance using Random Forest Method (link provided in reference section)

   - Visualize the top two columns (feature) in x-y coordinate system.
   - Remove the feature with the lowest importance and retrain your model using Decision Tree. Draw Decision and Evaluate performance of the model. Track the training time.
   - Remove the four features with the lowest importances and retrain your model using Decision Tree. Draw Decision tree and Evaluate performance of the model. Track the training time. - Remove the ten features with the lowest importances and retrain your model using Decision Tree. Draw Decision tree and Evaluate performance of the model. Track the training time.

7. (Analysis and Discussions) - Which model (from Q3 to Q6) performed the best?
   - Does removing least important features speed up training times?
   - Does removing least important features lower performance of your model?
   - How does removing less important features relevant to Big Data (extremely large dataset)?

Each Team will prepare a report. The report should include:

- Each question in "Coding" section and the **output** to the question.
- Analysis and discussions.
- References. A list of all references used. Example: links to websites where you found the codes. If you read some articles, please reference them. If you used Chatgpt, write down each question you asked.
- Files of your Source Code.

Each Team will submit the report and all codes related to the project.

Presentation (30 Points)

Each team is required give a presentation on their Project. During the presentation, each team will answer all the questions in the Coding section. Also helps to include your pseudo code. **Please be prepare to answer questions about your code and how it works.**

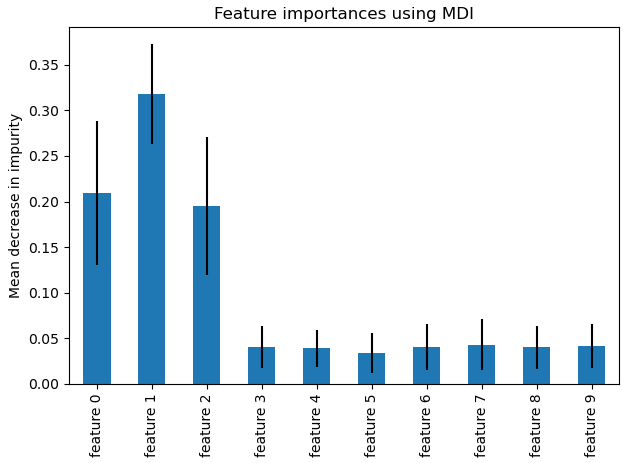

Figure 2: Example of Feature importance using Random Forest

Code Organization (7 points)

Please Organize your code. Your code should be inside Class(es) or function(s). Ex. If you reuse the same code twice, it would be useful to put those lines into a function. Your code should also include some comments, but it should not be excessive.

References

1. Dataset found here:
   https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

2. Link to sklearn
   https://scikit-learn.org/stable/

3. Dataset split:
   https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

4. Sample code for decision Tree:
   https://scikit-learn.org/stable/modules/tree.html

5. Sample code for Support Machine Vector
   https://scikit-learn.org/stable/modules/svm.html

6. Sample Feature importance code found here (using Random Forest)
   https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

7. Classification Performance measurement.
   https://scikit-learn.org/stable/modules/model_evaluation.html

8. Figure 1 reference: Shaymaa A. Hassan1, Mohammed S. Sayed1, , Mahmoud I. Abdalla1, Mohsen A. Rashwan, Detection of Breast Cancer Mass using MSER Detector and Features Matching