COSC 4P82: Assignment 1 - B
Using Genetic Programming in Diagnosis of Breast Cancer

Liam McDevitt
6015929

Andrew Pozzuoli
6017735

I. Wisconsin Breast Cancer Data Set

The Wisconsin breast cancer data set is a set of 569 instances of data describing X-ray images representing either a benign (B) or a malignant (M) tumor. The data set was received from the UCI Machine Learning Repository provided by the University of California School of Information and Computer Science [1]. The goal is to use genetic programming to evolve a function that takes the input and returns a diagnosis of malignant or benign that matches the actual diagnosis contained in the data set.

The problem is similar to a symbolic regression with a number of points input and output values where the output is malignant or benign and the input is the image values provided by the data set.

II. Experimental Setup

A. Parameters

TABLE I
GP Paramters

Parameter	Value
Population Size	500
Generations	40
Population Initialization	Ramped Half and Half
Minimum Grow Depth	2
Maximum Grow Depth	6
Selection	Tournament, k=4
Crossover	Subtree Crossover, 90%
Mutation	10%
Data Points Training	285
Data Points Testing	284
Runs Per Experiment	10

B. GP Language

The terminals used were thirty floating point values describing each X-ray image provided in the Wisconsin breast cancer date set.

Each value was repeated three times for mean, standard error, and largest.

C. Fitness Evaluation

The program evolved aimed to find a function that returned a negative value if the tumor was benign (B) or a positive value if the tumor was malignant (M).

TABLE II
Functions

Function	Description
Add	Adds two terminals.
Multiply	Multiplies two terminals.
Subtract	Subtracts two terminals.
Divide	Divides two terminals. If the denominator is zero then set the denominator to 1.
Negative	Makes the terminal negative.
Square Root	Takes the square root of a terminal. If the terminal is negative then take the square root of the absolute value of the terminal.
Exponent	Exponent with base e raised to the power of another terminal.
Log	Takes the natural logarithm of the absolute value of the terminal.

TABLE III
Terminals

Input	Description
Radius	Average radius of the tumor.
Texture	Gray-scale standard deviation.
Perimeter	Perimeter of the tumor.
Area	Area of the tumor.
Smoothness	Variation in radius.
Compactness	$\frac{perimeter^2}{area - 1.0}$
Concavity	Concave portion severity.
Concave Points	Number of concave portions.
Isymmetry	Isymmetry of the tumor
Fractal Dimension	coastline approximation - 1.0

The fitness was evaluated by measuring the number of correct diagnoses the function made (hits). Every time the evolved function produced a negative value and the input data corresponded to a benign diagnosis or if the function produced positive value and the input data corresponded to a malignant diagnosis, the sum of hits incremented by one. A higher hit count means a better fitness and zero hits would be the worst fitness.

The standardized fitness was calculated by taking the total number of diagnoses in the set and subtracting the number of hits. This gives the number of misdiagnoses (sum of errors), zero being the best and higher numbers being worse.

The adjusted fitness was calculated by subtracting the percentage of hits from one. This gives a number between 0 and 1 with 0 being the ideal and 1 being the worst.

III. Experiment

In our data set, there are 30 different image values we can use in our deciding factor on whether or not a patient has a malignant (M) or benign (B) tumor. Whether or not all 30 image values are important in our classification process is in question for our experiment.

For our experiment we are comparing the average fitness over 10 runs with two different sets of image values. The first set being the image values that were provided to us in the data set besides the patient identification number and, of course, the actual diagnosis. The second set being a small subset which contains only the radius image values from the data set. These include radius mean, standard error, and largest. The goal is to see if the performance of the GP is affected by reducing the number of image values besides the ones dealing with the radius of the tumor. We do note that there are many different possible combinations of image values that could perform better or worse depending on how these values work together. The experiment we are performing will only give us information on whether or not the radius of a tumor alone from an X-ray image is enough to accurately predict to a certain percentile whether it is malignant (M) or benign (B). However, the importance of this experiment is to see how the GP performs with different sets of image values for our classification problem.

For our experiment we partitioned the data set into a 50/50 split of training and testing data. Our fitness plots are based on how well the GP performed for the training sets. The 50/50 split of training and testing data is randomly shuffled each run to generate a good mix between our sets. In turn this should provide more general results on our testing set. To show the performance of our testing data we created Confusion Matrices. These Confusion Matrices are generated by using the best solution the GP found on the testing data set for each run. We will also be showing the best individual's tree we used for each Confusion Matrix.

IV. Results

A. Training Set Performance

Our adjusted fitness is used to plot our fitness in the range of [0, 1] with 1 being the worst and 0 being the best for visual purposes. This type of plot allows us to see convergence of our GP or premature convergence if it occurs. If premature convergence is occurring it is a good indication that GP does not need to be running for as many generations to achieve the same results.

The first fitness plot is showing the adjusted fitness of our GP running for 40 generations with all 30 image values as possible terminals in its calculation. As you can see from the graph, we have a steady decrease in the plotted output indicating that our GP is progressively getting better results as it evolves over the set number of generations for both the Population Average Fitness and the Best Fitness of Generation. However, you can see more of a difference in the average fitness in comparison to the best fitness in the rate of plotted decline. It is more important to look at the population's average fitness because it is a better indication on how well the GP is performing.

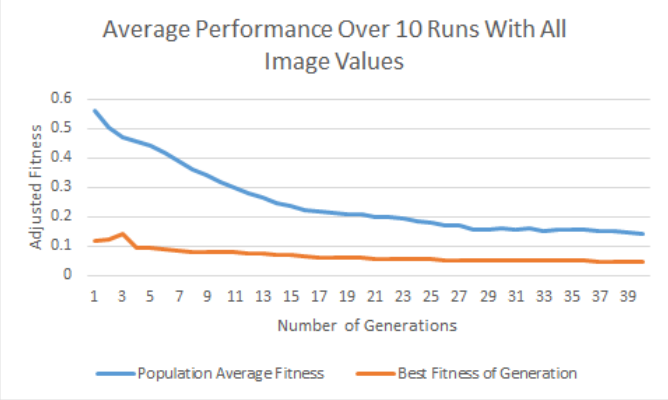

Fig. 1. Population average and average best performance over ten runs with all image values

The second fitness plot is showing the adjusted fitness of our GP running for 40 generations with only image values related to the radius of the tumor. These values include the radius mean, standard error, and largest. This fitness plot shows a steady improvement of the Population Average Fitness and the Best Fitness of Generations as the GP evolves over a set number of 40 generations. From the graph we can see the Best Fitness of Generation line converging at 94% of hits, meaning the GP is finding the correct result 94% of the time from the Best Fitness of Generation near the end of the run. The adjusted fitness is showing a steady improvement on the population's average fitness over the generations.

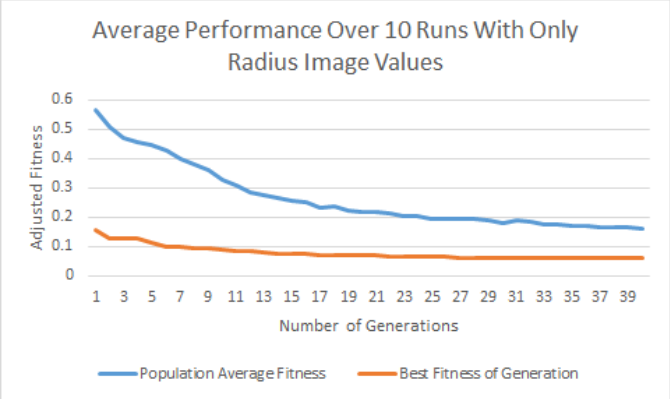

Fig. 2. Population average and average best performance over ten runs with only radius values (mean radius, radius standard error, largest radius

When looking at both fitness plots we can see they are very similar in both the Population Average Fitness and the Best Fitness of Generation values. The GP's performance on the training set was consistent between using all image values compared to using only the radius image values. This leads us to believe that the radius image values are very important in determining whether or not a tumor is malignant (M) or benign (B). Also, it is interesting to note that there appears to be little value in the other image values when it comes to making a correct decision. The similarity between the performances of these different sets of image values holds true on the training set but what about its tree's performance on data the GP has not seen before?

B. Testing Set Performance

For the Confusion Matrices we ran the two best runs' trees on our testing set to evaluate the GP's performance. We will show the tree corresponding the two best runs for each part of the experiment and Confusion Matrices for each tree applied to the testing set. If the fitness value of multiple runs were the same when trying to find the two best runs for each part of the experiment we took the one with the lowest number of hits in the false negatives because in regards to our problem we decided it would be better to tell a patient that their tumor is malignant (M) when it is actually benign (B) rather than telling the patient their tumor is benign (B) when it is actually malignant (M).

TABLE IV
CONFUSION MATRIX LEGEND

Legend		Actual Answer	
		M	B
GP	M	True Positives	False Positives
	B	False Negatives	True Negatives

```
Best Individual of Run:
Subpopulation 0:
Evaluated: true
Fitness: Standardized=8.0 Adjusted=0.1111111111111111 Hits=277
Tree 0:
 (* (sqrt (* (sqrt (/ smoothness_largest (/
     (/ smoothness_std_err perimeter_largest)
     concave_points_mean))) (0- (sqrt (/ (exp
     compactness_largest) (+ perimeter_std_err
     (* (exp area_std_err) (sqrt (/ (0- (0- radius_std_err))
         (/ (/ (/ area_mean perimeter_largest) concave_points_mean)
             (exp (ln (- (0- (exp concavity_largest))
                 (sqrt (0- texture_std_err))))))))))))))
     (* (sqrt (/ (exp concavity_largest) (/ (0-
         (ln (0- texture_mean))) (exp (/ (ln (exp
         (sqrt (ln area_mean)))) (- (* (+ (ln (- (0-
         (- texture_std_err concave_points_mean))
         (0- perimeter_largest))) (+ concavity_std_err
         radius_std_err)) (+ perimeter_std_err (*
         (exp area_std_err) (ln perimeter_largest))))
         (exp (exp isymmetry_std_err)))))))) (/ (/
         (0- (exp (* (sqrt texture_largest) (* (0-
             perimeter_largest) isymmetry_mean)))) (exp
         (exp concavity_largest))) (exp (sqrt (/ (exp
         (/ (ln (exp (/ (exp concavity_largest) (+
             (exp area_std_err) (exp concavity_largest)))))
             (- (* (+ texture_largest (+ concavity_std_err
                 radius_std_err)) (+ perimeter_std_err (*
                 (exp area_std_err) (ln perimeter_largest))))
                 (exp concavity_std_err)))) (/ (/ (+ (exp
         (sqrt (- smoothness_mean smoothness_largest)))
         (0- radius_std_err)) perimeter_largest) concave_points_mean)))))))
```

Fig. 3. Best individual of tens runs with all image values as possible terminals

TABLE V
ALL IMAGE VALUES TESTING SET BEST HITS

Hits		Actual Answer	
		M	B
GP	M	111	4
	B	15	154

TABLE VI
ALL IMAGE VALUES TESTING SET BEST PERCENTAGE HITS

% of Hits		Actual Answer	
		M	B
GP	M	97	2.5
	B	13	91

The best individual run on the testing set for all image values was 265 hits and 19 misses. Our tree structure is relatively complex from the best individual of the training set that we are using for this Confusion Matrix. We can see from the Confusion Matrix our training tree performed well on our testing data. Our true positives and true negatives were very high meaning the actual answer given by the data set matched the answer our GP gave by a high percentile. It is interesting to note that the number of false negatives is less than the number of false positives which is worse in our case for this specific problem.

```
Best Individual of Run:
Subpopulation 0:
Evaluated: true
Fitness: Standardized=9.0 Adjusted=0.1 Hits=276
Tree 0:
 (+ (* (sqrt (0- fractal_dimension_largest))
     (/ area_largest perimeter_largest)) (0- (sqrt
     (ln smoothness_std_err))))
```

Fig. 4. Second best individual of tens runs with all image values as possible terminals

TABLE VII
ALL IMAGE VALUES TESTING SET SECOND BEST HITS

Hits		Actual Answer	
		M	B
GP	M	95	15
	B	5	169

TABLE VIII
ALL IMAGE VALUES TESTING SET SECOND BEST PERCENTAGE HITS

% of Hits		Actual Answer	
		M	B
GP	M	86	8.6
	B	4.5	97

The second best individual run on the testing set for all image values was 264 hits and 20 misses. The tree structure for the best individual found by our training set is very small in comparison to the run before. In addition, the number of false negatives for this run is less than the number of false positives which is good for us in this case. If you take the difference in tree size and the number of false negatives in determining the performance of our GP this one would perform better for us. However, we are basing the performance on the number of hits which is out fitness and since it is one less it technically has a lower fitness than the one before in the definition of our performance measure of our GP.

TABLE IX
RADIUS IMAGE VALUES TESTING SET BEST HITS

Hits		Actual Answer	
		M	B
GP	M	86	10
	B	10	178

The best individual run on the testing set for only radius image values was 264 hits and 20 misses. By this fitness standard this one performed just all well and the second best performing run for all image values. The tree size for the best individual found by our training set is in-between the complexity of the first two run on all image values. Both of our true positives and true negatives are high which is good and out false negatives and false positives on the testing set were equivalent.

```
Best Individual of Run:
Subpopulation 0:
Evaluated: true
Fitness: Standardized=15.0 Adjusted=0.0625 Hits=270
Tree 0:
 (/ (ln (/ (* (exp (* (* (/ radius_std_err
     radius_mean) (sqrt (* (* (* (/ radius_std_err
     radius_mean) (sqrt (* radius_largest (sqrt
     radius_largest)))) (0- (* radius_largest
     radius_mean))) (ln radius_largest)))) (0-
     (* radius_largest radius_mean)))) (- (sqrt
     radius_std_err) (sqrt (+ (square (+ radius_std_err
     (0- radius_largest))) (square (* radius_largest
     (ln radius_largest)))))))) (exp radius_largest)))
     (exp (* (+ radius_largest (/ (ln (/ (ln (*
         radius_std_err (exp radius_largest))) (*
         radius_largest radius_mean))) (* (+ radius_largest
         radius_mean) (ln radius_largest)))) (square
         (- radius_largest radius_mean)))))
```

Fig. 5. Best individual of tens runs with radius image values as possible terminals

TABLE X
RADIUS IMAGE VALUES TESTING SET BEST PERCENTAGE HITS

% of Hits		Actual Answer	
		M	B
GP	M	90	5.3
	B	10.4	95

```
Best Individual of Run:
Subpopulation 0:
Evaluated: true
Fitness: Standardized=17.0 Adjusted=0.05555555555555555 Hits=268
Tree 0:
 (/ (/ (sqrt radius_std_err) (+ (- radius_std_err
     radius_mean) (exp (ln (- radius_largest (sqrt
     (- radius_std_err radius_mean)))))) (square
     (- radius_std_err (exp (sqrt (square (- (ln
         (- radius_largest (ln (- (sqrt radius_mean)
             radius_std_err)))) (* radius_largest (ln
         (exp (ln (- radius_largest radius_std_err))))))))))))
```

Fig. 6. Second best individual of tens runs with radius image values as possible terminals

TABLE XI
RADIUS IMAGE VALUES TESTING SET SECOND BEST HITS

Hits		Actual Answer	
		M	B
GP	M	87	17
	B	8	172

TABLE XII
RADIUS IMAGE VALUES TESTING SET SECOND BEST PERCENTAGE HITS

% of Hits		Actual Answer	
		M	B
GP	M	84	9.4
	B	7.7	96

The second best individual run on the testing set for only radius image values was 259 hits and 25 misses. This is the lowest performing one out of the best so far but it is still relatively close to the others. The true positive and true negatives are again in a very high percentile. The tree structure for the best individual found by our training set on this run is just a little bigger than the smallest one generated by the second best performing individual on the testing set for all image values. The number of false negatives is half of the number of false positives on our testing set which is good in our case based on alternate criteria thought of in comparing the same fitness as stated before.

V. Conclusion

From looking at how well our GP performed for both the training and testing sets for our experiment we can say not all image values are important for the GP's ability to determine whether or not a tumor is malignant (M) or benign (B) based on our data set. The experiment is not conclusive on which image values are important and which are not for our classification problem. However, we were able to show that the radius image values performed on par with all image values being used. A different subset could be made to maybe even perform better than our experiments subset chosen. This same experiment can be done by looking at many different combinations of image values but looking at every possible combination of image values and optimizing their performance would be a very long process. In doing so you could break this problem down to having the GP only looking at the absolute best subset of image values. Overall, this could lead to better performance on the GP for both training and testing sets. The conclusions and information on the testing set is far more important than the training because it shows how well our GP is performing on newly presented data. Our experiment showed that our best performing trees on the training set also performed well on the testing sets showing valuable ability by our GP to generalize accurate classifications

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.