COS710: Assignment 3

u22571532

April 2025

1 Dataset

The data set consists of variables representing medical attributes of a person with a target value depicting whether that person has hepatitis. The dataset is in *hepatitis.tsv*. The data is read into three corresponding arrays using the CSVReader.java with an array for the headers and target values each and a 2D array for the spec values.

Table 1: hepatitis.tsv

AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA
36.0	1	1	2	1	1	1
45.0	1	1	1	1	1	1

Table 2: Continued hepatitis.tsv

LIVER BIG	LIVER FIRM	SPLEEN PALPABLE	SPIDERS	ASCITES	VARICES
1	1	1	1	1	1
1	1	1	1	1	1

Table 3: Continued hepatitis.tsv

BILIRUBIN	ALK PHOSPHATE	SGOT	ALBUMIN	PROTIME	HISTOLOGY	TARGET
3	10	44	16	44	2	2
10	76	55	18	42	2	2

2 GP(Genetic Program) Algorithm

The GP makes use of a tree structure to make an individual of the population, representing a mathematical equation. The tree consists of two types of nodes, terminal and functional where functional nodes cannot be leaf nodes and terminal cannot be internal nodes.

- Terminal: represents the medical attribute values of the person well as constants ranging from 0 to 99
- Functional: represents the +, -, *, /, % and pow mathematical operators

The GP creates a population using given parameters : seed, epochs, population size, max tree depth, genetic operation split and data test & train split. During the run of the GP the best tree is stored to make sure the best solution found is not lost. The Genetic Programming algorithm implemented follows the evolutionary processes:

- **Preprocessing** : The records are read from the hepatitis.tsv file into three arrays. One is used to store the headers/variable names of the file excluding the target header. The second is a 2D array used to store the data, the columns correspond with the header array and the row correspond with the different records. The third is an array containing the target values of the records, corresponding with rows of the 2D arrays, duplicate records are allowed as overfitting is handled further in other ways. Although the dataset is very skew with most target values being 2, the dataset is not altered and an attempt to alleviate this bias is later discussed.

- **Initialization**: A random population is generated to fill the specified size. The population uses the grow method and therefore contains an even spread of trees of each depth up to the maximum depth with randomized fullness. The nodes are inserted at random to ensure great variation in the population. The root node is always a functional node.

- **Selection**: Tournament selection is used. Two random trees are chosen from the population and the one with the best fitness is chosen. Tournament selection is used to help nudge the population to have better fitness individuals with each generation. Tournament selection was chosen as it avoids elitism while still favoring better fitness.

- **Crossover**: Using two tournament selection calls, two parent trees are obtained. A random node is selected in the first parent tree. A second random node of the same type (function or terminal) is then selected from the second parent tree. The node of tree one is then placed in the spot of tree 2's chosen node and vice versa. The sub trees of the nodes stay with their parent node not changing, the same meaning they are also swapped between the two trees. The only nodes who's parent changes is the two chosen nodes. Deep copies of the trees are made to use for the next generation. The crossover event generates two individuals for the new population. This operator helps exploit the population to gain better results based on trees with already good results. Due to crossover being able to create bigger and bigger trees, a penalty was made to prevent trees from growing larger than to depths more than the specified max depth.

- **Mutation**: The mutation uses the selection function to obtain a singular tree. A random node of the chosen tree is selected. The chosen node is then mutated with a mutate function. If the node is a terminal node it is mutated it either a random other medical attribute or in a constant between 0 and 99. If it is a functional node it is mutated into a random operator of the ones stated earlier. Nodes stay the same type to avoid trees converging to an short tree local optima. A deep copy of the parent is used. This produces a one individual for the new generation.

- **Evaluation**: During training a fitness score is used this score does not represent the accuracy of the tree to obtain the score the following happens. The tree calculates value with the input(record) provide, this value is the run through a sigmoid function to get a value between 0 and 1. If the new value is smaller or equal to 0.5 it is classified as 1 (has hepatitis) otherwise it is classified as 2 (does not have hepatitis). If the calculated class matches the target class 2 points is added to the fitness total if the target class was 1 and 1 point is added if the target class was 2. If it target and calculated don't match the 2 points are deducted if the target value was 1 and nothing is added or deducted if the target value is 2. This done to try and improve accuracy of getting a correct "has hepatitis" with as minimal as

possible calculations saying ”no hepatitis” if the person ”has hepatitis”. This is important due to the bias the dataset has towards class 2 . The fitness score os then used during selection. Further in the fitness function 20 points are deducted from the fitness total if the trees depth exceeds the specified depth with 2 or is 2 shorter than that depth. This is done to prevent trees growing to unusable sizes.
To incorporate structure evaluation the following is done:
For the first half of the iterations global variability is promoted to explore the search space. This is done by calculating how similar each trees top three layers are to the rest of the population. The similarity score the tree to the rest of the population is then taken and subtracted from one. This value is then multiplied with the trees fitness to obtain it's new fitness. This means the lower it's similarity score is the less it's fitness changes. This is done to promote exploration of the global search space
In the second half of the generations local variation is promoted. The similarity score is calculated and use the same as above but it use the layers below layer three to calculate the similarity score. This is done to promote exploration in the local search space.
To conclude at the GP run the best tree derived is then run through an accuracy function to determine how accurately it predicts the target values of the testing set.

- **Termination**: The GP terminates after the specified epochs are reach. This is to prevent overfitting as well as to long run-times with minimal improvements. Further if the best tree solution does not change within 100 generations the GP terminates and that solution is provided.

- **Error Prevention**: The GP allows duplicates of the same tree to from and does not penalize this them. Leave nodes can not be function nodes, and their are checks in place to prevent and fix this. Non-leave nodes can not be terminal nodes, and their are checks in place to prevent and fix this. Mutations don't mutate nodes into different types to prevent the loss of entire trees. Crossovers only crossover the same type of node to prevent trees to shorten to the point of very non-optimal trees. Further null checks are present every where to prevent program crashes. Operators that have limitations like division and mod have checks to make sure they use a divisor of 0.

3 Parameters

The parameters used in the GP implementation look as follows:

Table 4: GP Default Parameters

Parameter	Value
Seed	Sys Time
Epochs	50
Population Size	50
Max Depth	4
Genetic Operator Split	0.5
Train/Test Split	0.8

Testing was done to determine the affect of each parameter (except the seed as it is always different for the tests). A variety of 5 different values were used for each parameter to see the affect. Each value ran 10 GP simulations. The average accuracy of each parameter value as well as the best run of that parameter value was saved. In the functions, the spec labels are place holders for their values in the dataset. A Genetic Operator Split of 0.3 indicates that 30% of the new population is made by mutation and 70% by crossover. Similarly, a Train/Test split of 0.8 shows that 80% of the dataset will be used for training and 20% for testing.

To find the most suitable parameters each parameter was changed five times while the other values parameters stayed the same as in Table 4. The results of those test were used to create a combination of parameters to test. Experimenting with these combinations lead to the parameters found in Table 5 that were used for the initial GP runs.

Table 5: GP Final Parameters for initial GP

Parameter	Value
Seed	Sys Time
Epochs	200
Population Size	50
Max Depth	6
Genetic Operator Split	0.4
Train/Test Split	0.8

As can be seen in Table 5 the max depth, 6, is quite big, this seems to be the case as their are quite a few variables/specs to be used in the function and this allows for greater utility. The mutation rate is also relatively on the lower side at 40% as the structure evaluation that promotes variability . The epochs are set as 200 as no increase in solution quality was constantly found with more generations. The population size of 50 was enough to gain great variability in the population. The epochs and population size were aimed to be kept small to save computational power and time, so that the max depth can be slightly increase without it greatly affecting the time it takes to find a solution.

4 Results

Table 6 contains the results of 10 consecutive NON - structure based GP runs using the parameters of Table 5.

Table 6: Results of 10 GP Runs

Run	Seed	Accuracy	Function
0	-892026561	0.2258	$(((((ASCITES \ast ASCITES) \ast ((VARICES - VARICES) - ((ALBUMIN \bmod ALBUMIN) \ast (SGOT/HISTOLOGY)))) \bmod ((VARICES - VARICES) \ast ((ALBUMIN - ALBUMIN) \ast (HISTOLOGY/HISTOLOGY)))) + (ANOREXIA \bmod 36.0)) \bmod ((VARICES/ALKPHOSPHATE) \ast (ALBUMIN - LIVERBIG)))$
1	-892025448	0.2177	$(((ASCITES \wedge SEX) - ((FATIGUE \ast (ASCITES/ANTIVIRALS))/(MALAISE - BILIRUBIN))) \ast ((MALAISE \ast (VARICES/STEROID)) - (ASCITES \ast BILIRUBIN)))$

Continued from previous page

Run	Seed	Accuracy	Function
2	-892024672	0.7016	$((((\mathit{LIVERFIRM} - \mathit{LIVERFIRM}) + ((\mathit{SPIDERS} + \mathit{ALKPHOSPHATE}) \ast \mathit{VARICES}))/(((\mathit{ASCITES} \ast \mathit{STEROID}) \bmod (((\mathit{LIVERFIRM} \bmod \mathit{LIVERFIRM}) + \mathit{ANTIVIRALS}) \ast ((21.0 \bmod \mathit{VARICES}) - (\mathit{VARICES} + \mathit{BILIRUBIN}))))/(((\mathit{SEX} \bmod \mathit{BILIRUBIN}) + \mathit{FATIGUE})/(\mathit{SPIDERS} - \mathit{SPIDERS})))) \bmod (((\mathit{SEX} \wedge \mathit{SPLEENPALPABLE}) \ast ((\mathit{STEROID} \bmod \mathit{BILIRUBIN}) - (\mathit{SEX} + \mathit{BILIRUBIN}))) - (((\mathit{LIVERFIRM} - \mathit{LIVERFIRM}) + ((\mathit{SPIDERS} + \mathit{SPIDERS}) \ast \mathit{VARICES}))/(((\mathit{ASCITES} \ast \mathit{SPIDERS}) \bmod (\mathit{MALAISE} \ast \mathit{PROTIME}))/(((\mathit{SEX} \bmod \mathit{BILIRUBIN}) + \mathit{FATIGUE}) - (\mathit{SPIDERS} - \mathit{SPIDERS})))))$

Continued from previous page

Run	Seed	Accuracy	Function
3	-891982719	0.8043	$((ASCITES \wedge BILIRUBIN) - (((((((ASCITES \wedge ANOREXIA) - ((((((((STEROID - ASCITES) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/MALAISE)) \wedge (ASCITES - SPIDERS)) - (((((((HISTOLOGY - SEX) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/STEROID)) + (ASCITES - LIVER\ BIG)) + (SGOT - ASCITES)) - (ALK\ PHOSPHATE \cdot SEX)) + ((((FATIGUE + ASCITES) + (FATIGUE - ANOREXIA))/((((HISTOLOGY - SPLEEN\ PALPABLE) + (ALK\ PHOSPHATE - SEX)) - (ALK\ PHOSPHATE \cdot ASCITES)) \bmod (ASCITES + ANTIVIRALS))) \bmod (HISTOLOGY/STEROID)))))/(SEX/SEX)) - (ASCITES/SPIDERS)) + ((((ASCITES \wedge ANOREXIA) - (((((SEX + SEX) - (ALK\ PHOSPHATE \cdot SEX)) + ((((FATIGUE + AGE) - (FATIGUE - ANOREXIA))/((((((((HISTOLOGY - SEX) - (ALK\ PHOSPHATE - SEX))/((ASCITES \wedge STEROID)/(((HISTOLOGY \bmod STEROID) - (((((((HISTOLOGY + SPLEEN\ PALPABLE) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/ANOREXIA)) + (ASCITES - SPIDERS)) + (ALK\ PHOSPHATE - SEX)) - (ALK\ PHOSPHATE - ASCITES))/(ANOREXIA - ANOREXIA))) + ((ASCITES \wedge BILIRUBIN)/(SEX \bmod SEX))))) + (VARICES - LIVER\ BIG)) + (SGOT - ASCITES)) + (ALK\ PHOSPHATE \cdot SEX)) - (ALK\ PHOSPHATE \cdot ASCITES)) \bmod (ASCITES + ANTIVIRALS))) \bmod (HISTOLOGY/STEROID))) - ((HISTOLOGY - SEX) \cdot (ALK\ PHOSPHATE - SEX))) + ((((ANOREXIA + STEROID) - (ALK\ PHOSPHATE - ANOREXIA))/(HISTOLOGY - STEROID)) \bmod (HISTOLOGY/STEROID))))/(SEX/ (HISTOLOGY/STEROID)))))/(SEX/SEX)) - (ASCITES/SPIDERS)) + (ALK\ PHOSPHATE - SPLEEN\ PALPABLE)) - (ALK\ PHOSPHATE - ASCITES))/(HISTOLOGY + LIVER\ BIG))$

Continued from previous page

Run	Seed	Accuracy	Function
4	-892021741	0.2500	$((((ALBUMIN \wedge ALBUMIN) * (HISTOLOGY - (6.0 - STEROID))) \bmod ((ALBUMIN \wedge ALBUMIN)/((LIVERBIG + LIVERBIG)/MALAISE)))/((((SPLEENPALPABLE\ AGE) \wedge (VARICES * VARICES)) - ((ASCITES \bmod ALKPHOSPHATE)/((ASCITES * MALAISE)/FATIGUE))) + (((AGE + SPLEENPALPABLE) - (BILIRUBIN * BILIRUBIN)) \wedge SEX)))$
5	-892020994	0.2177	$(((((((VARICES/BILIRUBIN) * ASCITES) \wedge (PROTIME + (ALBUMIN/MALAISE))) - ((ANTIVIRALS * (ANTIVIRALS + SPIDERS)) - (STEROID * STEROID))) - ((ANOREXIA/PROTIME)/ASCITES)) * ((HISTOLOGY * (SPIDERS + STEROID)) - (STEROID * MALAISE))) - (((LIVERFIRM + (ALBUMIN - MALAISE)) + (SPIDERS \bmod MALAISE))/(((HISTOLOGY + ANOREXIA) \bmod (ALKPHOSPHATE + AGE)) \bmod (SGOT/BILIRUBIN))))$
6	-892019484	0.2419	$((((SPIDERS + ASCITES) \bmod ((SPIDERS * ASCITES) * (VARICES/MALAISE))) \wedge ((SGOT - (STEROID * MALAISE)) + ((SPIDERS - VARICES) + ((VARICES + MALAISE) - (MALAISE + SPIDERS))))) \bmod ((((BILIRUBIN - BILIRUBIN) \bmod MALAISE)/((VARICES/MALAISE) * (BILIRUBIN \wedge BILIRUBIN)))/(((SEX - MALAISE) + MALAISE) + (LIVERFIRM * ASCITES))))$
7	-892017680	0.3871	$((((SEX \wedge VARICES) \bmod ((LIVERBIG - ASCITES) * (VARICES/VARICES))) \bmod (((AGE * MALAISE) * (PROTIME \bmod LIVERFIRM)) + ((SEX \bmod VARICES)/((MALAISE \bmod MALAISE) \bmod (FATIGUE - MALAISE))))) + (((LIVERBIG - ASCITES) * (VARICES \bmod VARICES)) + ((ASCITES * MALAISE)/(VARICES \bmod VARICES))))$
8	-892016243	0.3387	$(((ANOREXIA \bmod VARICES) * FATIGUE) \wedge ((LIVERBIG + LIVERBIG) \bmod ALBUMIN))$

Continued from previous page

Run	Seed	Accuracy	Function
9	-892015282	0.3306	$(((ASCITES \bmod BILIRUBIN)/((ANOREXIA * STEROID)/(((VARICES \wedge STEROID) \bmod (MALAISE - ANOREXIA)) \bmod ((PROTIME \bmod SEX) - (BILIRUBIN + SPIDERS)))))/(((SEX * LIVERBIG) + ((((SEX \wedge STEROID) + (LIVERFIRM - ANOREXIA)) \bmod (BILIRUBIN - MALAISE)) \bmod (BILIRUBIN - MALAISE))) * ((((VARICES \wedge STEROID) \bmod (MALAISE - ANOREXIA)) \wedge ((PROTIME \bmod SEX) - (BILIRUBIN + SPIDERS))) - (HISTOLOGY/MALAISE))))$
	Average Accuracy		0.3613
	Standard Deviation		0.1789
	Best Seed		-891982719
	Best Accuracy		0.8043

Continued from previous page

Run	Seed	Accuracy	Function
	Best Solution Equation		$((ASCITES \wedge BILIRUBIN) - (((((((ASCITES \wedge ANOREXIA) - ((((((((STEROID - ASCITES) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/MALAISE)) \wedge (ASCITES - SPIDERS)) - (((((((HISTOLOGY - SEX) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/STEROID)) + (ASCITES - LIVER\ BIG)) + (SGOT - ASCITES)) - (ALK\ PHOSPHATE \cdot SEX)) + ((((FATIGUE + ASCITES) + (FATIGUE - ANOREXIA))/((((HISTOLOGY - SPLEEN\ PALPABLE) + (ALK\ PHOSPHATE - SEX)) - (ALK\ PHOSPHATE \cdot ASCITES)) \bmod (ASCITES + ANTIVIRALS))) \bmod (HISTOLOGY/STEROID)))))/(SEX/SEX)) - (ASCITES/SPIDERS)) + ((((ASCITES \wedge ANOREXIA) - (((((SEX + SEX) - (ALK\ PHOSPHATE \cdot SEX)) + ((((FATIGUE + AGE) - (FATIGUE - ANOREXIA))/((((((((HISTOLOGY - SEX) - (ALK\ PHOSPHATE - SEX))/((ASCITES \wedge STEROID)/(((HISTOLOGY \bmod STEROID) - (((((((HISTOLOGY + SPLEEN\ PALPABLE) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/ANOREXIA)) + (ASCITES - SPIDERS)) + (ALK\ PHOSPHATE - SEX)) - (ALK\ PHOSPHATE - ASCITES))/(ANOREXIA - ANOREXIA))) + ((ASCITES \wedge BILIRUBIN)/(SEX \bmod SEX))))) + (VARICES - LIVER\ BIG)) + (SGOT - ASCITES)) + (ALK\ PHOSPHATE \cdot SEX)) - (ALK\ PHOSPHATE \cdot ASCITES)) \bmod (ASCITES + ANTIVIRALS))) \bmod (HISTOLOGY/STEROID))) - ((HISTOLOGY - SEX) \cdot (ALK\ PHOSPHATE - SEX))) + ((((ANOREXIA + STEROID) - (ALK\ PHOSPHATE - ANOREXIA))/(HISTOLOGY - STEROID)) \bmod (HISTOLOGY/STEROID))))/(SEX/$ $(HISTOLOGY/STEROID)))))/(SEX/SEX)) - (ASCITES/SPIDERS)) + (ALK\ PHOSPHATE - SPLEEN\ PALPABLE)) - (ALK\ PHOSPHATE - ASCITES))/(HISTOLOGY + LIVER\ BIG))$

Continued from previous page

Run	Seed	Accuracy	Function

After the final tests we saw that the Accuracy of the non - structure based GP was 0.3613 . The standard deviation was found to be 0.1789, meaning the GP has some variation in accuracy depending seed used.

The best solution found as represented in Table 7 had an Accuracy of 0.8043 meaning on average the solution derived has a 80.43% to be the correct classification.

Table 7: The best result achieved in the 10 NON - structure based GP runs

Seed	Accuracy	Function
-891982719	0.8043	$((ASCITES \wedge BILIRUBIN) - ((((((ASCITES \wedge ANOREXIA) - (((((((STEROID - ASCITES) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/MALAISE)) \wedge (ASCITES - SPIDERS)) - (((((((HISTOLOGY - SEX) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/STEROID)) + (ASCITES - LIVER\ BIG)) + (SGOT - ASCITES)) - (ALK\ PHOSPHATE \cdot SEX)) + ((((FATIGUE + ASCITES) + (FATIGUE - ANOREXIA))/((((HISTOLOGY - SPLEEN\ PALPABLE) + (ALK\ PHOSPHATE - SEX)) - (ALK\ PHOSPHATE \cdot ASCITES)) \bmod (ASCITES + ANTIVIRALS))) \bmod (HISTOLOGY/STEROID)))))/(SEX/SEX)) - (ASCITES/SPIDERS)) + ((((ASCITES \wedge ANOREXIA) - (((((SEX + SEX) - (ALK\ PHOSPHATE \cdot SEX)) + ((((FATIGUE + AGE) - (FATIGUE - ANOREXIA))/((((((((HISTOLOGY - SEX) - (ALK\ PHOSPHATE - SEX))/((ASCITES \wedge STEROID)/(((HISTOLOGY \bmod STEROID) - (((((((HISTOLOGY + SPLEEN\ PALPABLE) - (ALK\ PHOSPHATE - SEX))/(HISTOLOGY/ANOREXIA)) + (ASCITES - SPIDERS)) + (ALK\ PHOSPHATE - SEX)) - (ALK\ PHOSPHATE - ASCITES))/(ANOREXIA - ANOREXIA))) + ((ASCITES \wedge BILIRUBIN)/(SEX \bmod SEX))))) + (VARICES - LIVER\ BIG)) + (SGOT - ASCITES)) + (ALK\ PHOSPHATE \cdot SEX)) - (ALK\ PHOSPHATE \cdot ASCITES)) \bmod (ASCITES + ANTIVIRALS))) \bmod (HISTOLOGY/STEROID))) - ((HISTOLOGY - SEX) \cdot (ALK\ PHOSPHATE - SEX))) + ((((ANOREXIA + STEROID) - (ALK\ PHOSPHATE - ANOREXIA))/(HISTOLOGY - STEROID)) \bmod (HISTOLOGY/STEROID))))/(SEX/SEX)) \bmod (HISTOLOGY/STEROID))))/(SEX/SEX)) - (ASCITES/SPIDERS)) + (ALK\ PHOSPHATE - SPLEEN\ PALPABLE)) - (ALK\ PHOSPHATE - ASCITES))/(HISTOLOGY + LIVER\ BIG))$

Table 8: Results of 10 structure based GP Runs

Run	Seed	Accuracy	Function
0	-387499898	0.3065	$((((SPIDERS + FATIGUE) * (SPIDERS \bmod (BILIRUBIN \bmod VARICES)))/((ALBUMIN - ANOREXIA) \wedge (BILIRUBIN - PROTIME))) - (((((SPIDERS + ASCITES) * (SPIDERS \bmod (ALKPHOSPHATE \wedge PROTIME))) + (SEX \wedge LIVERFIRM)) + (SPIDERS + (BILIRUBIN + VARICES))) * (((HISTOLOGY \bmod ANOREXIA) \bmod (SPIDERS - ASCITES)) + LIVERBIG)))$
1	-387498565	0.8226	$((((VARICES + (((MALAISE * ASCITES) * BILIRUBIN) * BILIRUBIN)) - (ANTIVIRALS \bmod HISTOLOGY)) - ((((MALAISE \bmod ASCITES)/BILIRUBIN) - ((ANTIVIRALS/HISTOLOGY) \wedge (SPIDERS + FATIGUE))) \wedge (((MALAISE + ASCITES)/BILIRUBIN) * ((STEROID \bmod SPIDERS) - SPIDERS)))) + ((ANTIVIRALS/HISTOLOGY) \wedge (SPIDERS + ASCITES)))$
2	-387497161	0.8709	$(((VARICES/ALKPHOSPHATE)/((STEROID/FATIG$ $MALAISE)) - (((ASCITES \bmod ALKPHOSPHATE) * ((STEROID - STEROID) - SEX))/((15.0 \wedge LIVERFIRM) - ((VARICES - ANOREXIA) \bmod (((AGE + AGE)/(HISTOLOGY/AGE)) * MALAISE)))))$
3	-387496153	0.8226	$((((ANTIVIRALS \bmod BILIRUBIN) + ((ALBUMIN \bmod VARICES)/((SPIDERS * MALAISE) + (VARICES * BILIRUBIN)))) * ((ALBUMIN \bmod VARICES) + ((41.0 + MALAISE) + (FATIGUE/BILIRUBIN)))) * (((((PROTIME + ANTIVIRALS) \wedge (ANTIVIRALS/PROTIME)) * ASCITES) - ((SPIDERS * MALAISE)/(ANTIVIRALS + BILIRUBIN))) \bmod (((2.0 - SPLEENPALPABLE) + (SEX + ANOREXIA)) \wedge (((ANTIVIRALS - VARICES) + ((VARICES + FATIGUE) * (SPIDERS/ALBUMIN))) \bmod SEX))))$
4	-387494122	0.2177	$(((SPLEENPALPABLE/SPLEENPALPABLE) \wedge ((SPLEENPALPABLE + PROTIME) \wedge (STEROID - (PROTIME - PROTIME)))) \bmod (SEX \bmod MALAISE))$

Continued from previous page

Run	Seed	Accuracy	Function
5	-387493458	0.7823	$((((SGOT \wedge (AGE - LIVERBIG))/((STEROID+LIVERFIRM)-(ANOREXIA * STEROID))) \bmod (((ANOREXIA + AGE) \wedge (BILIRUBIN \wedge BILIRUBIN)) - ((ALKPHOSPHATE * SGOT) - SGOT))) - (STEROID \bmod FATIGUE))$
6	-387492784	0.6855	$(((((MALAISE + MALAISE) \wedge SEX) + ((LIVERBIG/LIVERBIG) * LIVERBIG))/(LIVERBIG \bmod AGE)) \bmod (((ALBUMIN \wedge (SEX \bmod SEX))/((FATIGUE/FATIGUE) \bmod (FATIGUE * FATIGUE))) * (((SPLEENPALPABLE/SPLEENPALPABLE)-(2.0 + LIVERFIRM)) - ((SPIDERS + SPIDERS) \wedge SPIDERS))))$
7	-387492040	0.2258	$((((((STEROID * STEROID) * FATIGUE) \bmod (ASCITES/((MALAISE * ANTIVIRALS) \bmod BILIRUBIN)))*((AGE*ANTIVIRALS)*(STEROID-STEROID)))\wedge ((ALKPHOSPHATE*ALKPHOSPHATE)\wedge (ASCITES * ((MALAISE \bmod BILIRUBIN) \bmod BILIRUBIN)))) \bmod (((SPIDERS/ALKPHOSPHATE)/(BILIRUBIN+(ANOREXIA * ALBUMIN))) + (ASCITES/((55.0 \bmod SPLEENPALPABLE) + ALBUMIN))))$
8	-387491000	0.3548	$(((((LIVERFIRM + STEROID)/(((FATIGUE + SEX) + (SPIDERS * MALAISE)) * (SPIDERS - 28.0)))/(((SPLEENPALPABLE/LIVERBIG)/((SPIDE VARICES)*(ANTIVIRALS/ALBUMIN)))/(LIVERFIR ((((ANOREXIA + STEROID) + (ALKPHOSPHATE * MALAISE)) \bmod ((BILIRUBIN/MALAISE) - ((SPIDERS * SPIDERS) * (ASCITES/ALBUMIN)))) * (((SPIDERS/STEROID) * ((SPIDERS * VARICES) * (ASCITES * ALBUMIN))) * (SPLEENPALPABLE + LIVERBIG)))) \wedge (LIVERBIG/LIVERBIG))$
9	-387489533	0.7823	$((((ASCITES + ALKPHOSPHATE) * (ALBUMIN/ALKPHOSPHATE)) \bmod (BILIRUBIN + (ALBUMIN \bmod FATIGUE)))-((HISTOLOGY-SPIDERS)*((ANTIVIRALS \bmod ANTIVIRALS) \wedge STEROID)))$
	Average Accuracy		0.5847
	Standard Deviation		0.2576

Continued from previous page

Run	Seed	Accuracy	Function
	Best Seed		-387497161
	Best Accuracy		0.8709
	Best Solution Equation		$(((VARICES/ALKPHOSPHATE)/((STEROID/FATIG$ $MALAISE)) - (((ASCITES \bmod ALKPHOSPHATE) * ((STEROID - STEROID) - SEX))/((15.0 \wedge LIVERFIRM) - ((VARICES - ANOREXIA) \bmod (((AGE + AGE)/(HISTOLOGY/AGE)) * MALAISE)))))$

After the final tests we saw that the Accuracy of the runs is 0.5847 . The standard deviation was found to be 0.2576, meaning the GP has some variation in accuracy depending seed used.

The best solution found as represented in Table 9 had an Accuracy of 0.8709 meaning on average the solution derived has a 87.09% to be the correct classification.

Table 9: The best result achieved in the 10 structure based GP runs

Seed	Accuracy	Function
-387497161	0.8709	$(((VARICES/ALKPHOSPHATE)/((STEROID/FATIGUE) \bmod$ $MALAISE)) - (((ASCITES \bmod ALKPHOSPHATE) * ((STEROID - STEROID) - SEX))/((15.0 \wedge LIVERFIRM) - ((VARICES - ANOREXIA) \bmod (((AGE + AGE)/(HISTOLOGY/AGE)) * MALAISE)))))$

5 Conclusion

Due to the imbalanced dataset set it was difficult to create a GP that was not biased towards always providing the common class just because it will have a good accuracy. Thus accuracy was not used for the fitness, but more complicated fitness calculation method was used ,as described earlier, to try and combat this. Further, we saw that using a structured based GP also promoted the finding of better results as it promoted variability in the population in the global and local search space. This was good as the dataset caused this GP to struggle to try and predict the classification when it when had to get a worse fitness before getting a better one. The tree pruning (reducing fitness if the depths became to large or to small) was also very help full as it kept the trees from becoming to large as well as to small. Overall, the use of structure control helped improve the average accuracies of the GP. We see this as with non structure control the the average accuracy was 36.13% and with structure control it had a 84.68% average accuracy. The use of global and local variability enforcement was very effective as it expanded the search space where needed with it constantly removing individuals with a good fitness but low variability.