HYBRIDIZATION OF MULTILAYER PERCEPTRON AND GENETIC ALGORITHM FOR LUNG CANCER DISEASE DIAGNOSIS USING MICRO-ARRAY DATASET

BY

¹Ahmed Abiodun Taofik, ²Issah Abolaji Yusuf & ³Muhammed Kamaldeen Jimoh ^{1&2}Department of Computer Science, School of ICT and Management Science, Kwara State College of Arabic and Islamic Legal Studies, Ilorin. ³Department of Educational Technology, University of Ilorin, Ilorin, Nigeria. Email: ahmedtaophic1@gmail.com

Abstract

This paper attempt a shift in paradigm from conventional methods by formulating hybridizing model of genetic algorithm and Multilayer perceptron for optimization of relevant features of the genes and classification of lung cancer disease respectively. Microarray data is to be considered as a dataset. The paper therefore, adopt hybridize model of Genetic Algorithm and MLP, it was developed and simulated in Weka environment using microarray cancer dataset. The solution found by the combined Genetic Algorithm and Multilayer Perceptron performed effectively well. The results presented in this paper revealed that the proposed hybridization of Genetic Algorithm and Multilayer Perceptron performs better with over 90% accuracy when used for classification of microarray dataset of lung cancer. **Keywords:** Cancer, Diagnosis, Genetic Algorithm, Multilayer Perceptron, Hybridization

Introduction

Early detection of cancer disease is the key of its cure. The automatic diagnosis of cancer is an important, real-world medical problem. Cancer is one of the most common and deadly diseases in the world Ganesan, et al., (2020). The conventional diagnostic techniques are not always effective as they rely on the physical and morphological appearance of the tumor. According to Khalid and Atif (2020), early stage prediction and diagnosis is difficult with those conventional techniques. Moreover, these techniques are also costly, time consuming, requires large laboratory setup and highly skilled persons. It is well known that cancers are involved in genome level changes. Thus, it implies that for a specific type of cancer there could be pattern of genomic change. If those patterns are known, then it can serve as a model for the detection of that cancer and will help in making better therapeutic decisions Singh, et al., (2019) as quoted in Khalid and Atif (2020). Cancer is one of the most common deadly diseases in the world. The conventional diagnostic techniques are not always effective as they rely on the physical and morphological appearance of the tumor (Jimoh, 2015). Artificial intelligence is a branch of computer science and a discipline in the study of machine intelligence that is, developing intelligent machines or intelligent systems imitating, extending and augmenting human intelligence through artificial means and techniques to realize intelligent behavior. Its techniques offer advantages such as adaptation, fault tolerance, learning and human-like behavior over conventional computing techniques. The idea is to combine the pathological, intelligent and statistical approaches to enable simple and accurate diagnosis and prognosis. Artificial intelligence has been used in various areas such as cancer diseases diagnosis. It was revealed by American Cancer Society (2020) that cancer begins when cells in a part of the body start to grow out of control. There are many kinds of cancer, but they all start because of out-of-control growth of abnormal cells. Cancer cell growth is different from normal cell growth. Instead of dying, cancer cells continue to grow and form new abnormal cells. Cancer cells can also invade (grow into) other tissues, something that normal cells cannot do. Growing out of control and invading other tissues are what makes a cell a cancer cell American Cancer Society (2020).

The ability of the physicians to effectively treat and cure cancer is directly dependent on their ability to detect cancers at their earliest stages. According to World Health Organization –WHO (2021), cancer is a generic term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumours and neoplasms. Cancer is a leading cause of death worldwide, accounting for about 10 million deaths in 2020. The most common causes of cancer death are cancers of lung (about 3 million deaths), liver (about 1 million deaths), stomach (1 million deaths), colorectal (about 1 million deaths), breast (more than 521 000 deaths) and oesophageal cancer (more than 400 000 deaths) (WHO, 2020). Projections based on the GLOBOCAN 2020

The scope of this research is to apply Multilayer Perceptron and genetic algorithms techniques to Health care, specifically to the diagnosis of lung cancer patients. Also, it involves formulating and implementing a neural-genetic crossbreeding model to develop a system for diagnosing lung cancer disease using microarray data. A comprehensive study of the process of neural network such as (learning process, transfer function, back-propagation algorithm, feed-forward networks, network layers, perceptron, selection of weights, data description and training of data) and main ingredients of genetic algorithm such as chromosome, fitness, selection and crossover / Mutation were explored and implemented.

Literature Review

According to American Canser Society (2021), cancer can be described as a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Cancer is caused by both external factors (tobacco, infectious organisms, chemicals, and radiation) and internal factors (inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These causal factors may act together or in sequence to initiate or promote the development of cancer. Ten or more years often pass between exposure to external factors and detectable cancer. According to Cancer Research UK (2020), bodies receive oxygen through the lungs and pass it into the bloodstream so that it can circulate to everybody cell. The muscles of our chest and a large flat muscle under the lungs (the diaphragm - pronounced di-a-fram) are used to draw air into the lungs. The diaphragm is at the base of the chest cavity, just above the stomach. The chest cavity is sealed so that when you breathe in and the muscles make it bigger, this creates a vacuum inside, which draws air in through your nose and down into the lungs. Khalid and Atif (2020) presented a comprehensive evaluation of machine learning techniques for cancer class prediction based on microarray data, various techniques were implied on prostate cancer dataset in oder to accuratly predict cancer class. The reserchers applied combination of statistical techniques such as inter-quartile range and t-test, which has been effective in fitering significant genes and minimizing noies from data. However, each technique were handle monolthically on the prostate cancer dataset and this approch does not uselung cancer dataset. Shelly, et al., (2011) conducted a survey on various data mining classification techniques for enhancing breast cancer diagnosis and prognosis. The reserachers also summarized various related articles on breast cancer diagnosis and prognosis. However, this work does not consider development of an enhanced approach for lung cancer diagnosis. Vitoantonio, et al., (2006) proprosed approch of combining two techniques which are genetic algorithms and atificial neural networks to analyse microarray data as a distributed approach. In the research work, the researchers address the problem of gene selection using a distributed genetic algorithm that evolves populations of possible solution and uses an artificial neural network in order to test the gane signatures and ability to correctly classify cases belonging to the test set. The researchers did not apply these techniques to solve the problem of cancer disease diagnosis

Genetic algorithms (GAs) "were invented by John Holland in the 1960s and were developed by Holland and his students and colleagues at the University of Michigan in the 1960s and the 1970s. In contrast with evolution strategies and evolutionary programming, Holland's original goal was not to design algorithms to solve specific problems, but

rather to formally study the phenomenon of adaptation as it occurs in nature and to develop ways in which the mechanisms of natural adaptation might be imported into computer systems. Multilayer Perceptron have had a unique history in the realm of technology. Unlike many technologies today which either immediately fail or are immediately popular, neural networks for popular for a short time, took a two-decade hiatus, and have been popular ever since (Eric, 2020). The first step toward artificial neural networks came in 1943 when Warren McCulloch, a neurophysiologist, and a young mathematician, Walter Pitts, wrote a paper on how neurons might work. They modeled a simple neural network with electrical circuits. Reinforcing this concept of neurons and how they work was a book written by Donald Hebb. Every cell in the human body carries an individual's genetic information in DNA, of which genes are specific parts that encode for proteins to allow biological activity to occur. Whether certain genes are active or not can be measured using microarrays, which can probe tens of thousands of genes simultaneously (Tom, 2005). The first arrays, created in the mid80s, were called macro arrays. They were fabricated by spotting DNA probes on a membrane-type material with spot sizes of about 300 microns, which limited the density of the spots to about 2000 probes.

Research Objectives

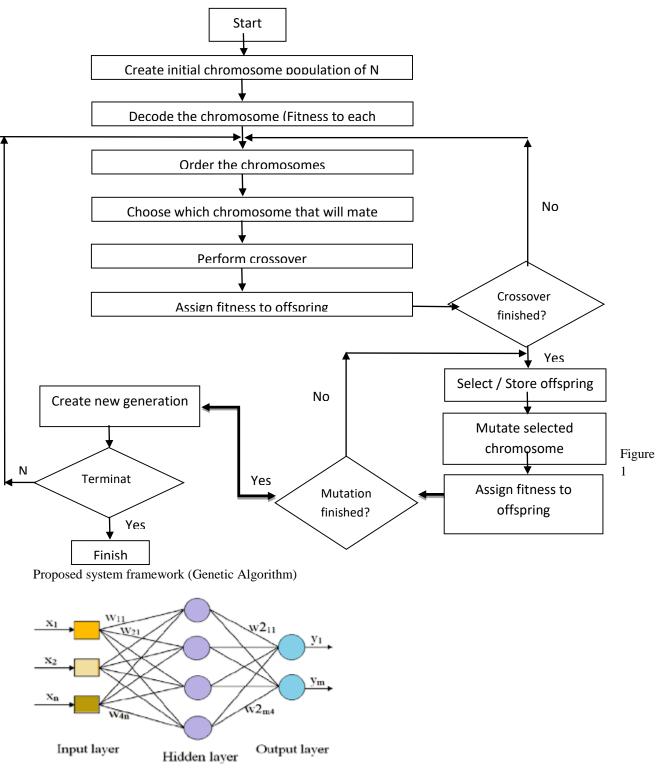
The objectives are:

- i. to formulate genetic- multilayer perceptron for lung cancer disease diagnosis;
- ii. to implement the formulated model in WEKA (Waikato Environment for Knowledge Analysis) development environment;
- iii. to validate the efficiency of the model
- iv. to know he accuracy of the model

Methodology

Proposed System Framework

The proposed system consists of different modules and divided into two stages as shown in Figure 1 and 2 Stage one describe how the complexity of microarray data is reduced using Genetic Algorithm while stage two involves the use of Neural Network for cancer diseases classification.



Stage 1: Genetic Algorithms to reduce the complexity of the microarray data

. (IJARMS) 165

Figure 2 Proposed system frame work (Multilayer Perceptron)

The above process is repeated until some condition is satisfied (Rahul, Narinder, and Yaduvir, (2019)). Algorithmically, the basic genetic algorithm (GAs) is outlined as below:

Step 1: Generate initial Population P(0) at random, and set i=0;
Step 2: repeat
Step 3: Evaluate the fitness of each individual in P(i)
Step 4: Select parents from P(i) based on their fitness in P(i)
Step 5: Apply crossover to create offspring from parents
Step 6: Apply mutation to the offspring
Step 7: Select generation P(i+1) from current offspring, O(i), and parents P(i)
Step 8: Until finished

Findings and Discussion

The goal of this research is to hybridize genetic algorithm and Multilayer Perceptron for lung cancer disease diagnosis based on examine data. Genetic algorithm reduces complexity of microarray dataset, that is, feature selection of the most relevant attributes and MLP does the classification. The implementation was carried out using Weka (Waikato Environment for Knowledge Analysis) which was developed at the University of Waikato, New Zealand. Weka is written in Java and commonly used for machine learning. Its features include preprocessing, classification, clustering, association, feature selection and visualization among others. The experiments were carried out on a 64-bit operating system with Windows 8.0, Intel(R) Core(Tm) i7- 3632QM CPU @ 2.20GHz and 8Gb of RAM. Due to the iterative nature of the experiments and resultant processing power required, the Java heap size for Weka version 3.6.12 was set to 1024MB to assess the effectiveness of the algorithms. The dataset used in this work is lung cancer dataset and was taken from biomedical dataset repository (http://www.chestsurg.org/microarray.htm) for the classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 149 tissue samples (15 MPM and 134 ADCA). The training set contains 69 (46.31%) of them, 10 MPM and 59 ADCA. The rest 80 (53.69%) samples are used for testing as shown in figure 3. (a and b) and 4 (a and b) respectively. Each sample is described by 12533 genes and figure 5 shows the visualization of the microarray dataset. Figure 6 show the spreadsheet copy of both training and testing dataset respectively.



Figure 3(a): The training set contains 69 of tissue samples, 10 MPM and 59 ADCA

Instances: 6 Itributes			 12534	
conducters.				
	All	None	Invert	Pattern
ю.	Name			
A al DUPP	TAON TO THE			
12505	968_J_at			
12506	969_#_#t			
12507	970_r_at			
12508	971_s_at			
12509	973_at			
12510	974_at			
12511	975_at			
12512	976_9_at			
12513	977_s_at			
12514	978_at			
12515	979_g_at			
12516	980_at			
12517	981_at			
12510	982_at			
12519	983_at			
12520	984_g_at			
12521	986_at			
12522	987_g_at			
12523	988_at			
12524	989_at			
12525	990_at			
12526	991_g_at			
12527	992_at			
12528	993_at			
12529	994_at			
12530	995_g_at			
12531	996_at			
12532	998_s_at			
12533	999_at			

Figure 3 (b): The training set contains 59 of tissue samples and each sample is described by 12533 attributes plus class attribute making total of 12534 attributes

Selected att Name: C Missing: 0	lass	Distinct:	2		Type: No Unique: 0	ominal (0%)	
No.	Label			Count	t		
1	Mesothelioma			5			
2	ADCA			75			
Class: Class	(Nom)				~	Visualize	All
			75				
5		_	_	_	Log		►. × 0

Figure 4(a): The testing set contains 80 of tissue samples,5 MPM and 75 ADCA

Attribute	es					
	All		None	Invert	Pattern	
No.		Name	_			
12	522	987_g	at			~
12	523	988_a	t			1
12	524	989_a	t			1
12	525	990_a	t			1
12	526	991_g	_at			1
12	527	992_a	t			1
12	528	993_a	t			1
12	529	994_a	t			1
12	530	995 <u>g</u>	_at			1
12	531	996_a	t			1
12	532	998_s	_at			1
12	533	999_a	t			1
12	534	 Class 				\sim

Figure 4(b): The testing set contains 149 of tissue samples and each sample is described by 12533 attributes plus class attribute making total of 12534 attributes

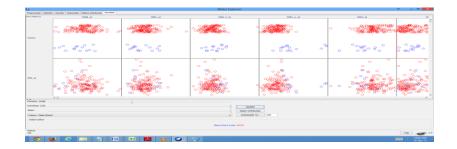


Figure 5: The visualization of the microarray dataset

								viewer								
elatic		d by huigir														
No.	1000_at Numeric	1001_at Numeric	1002_f_at Numeric	1003_s_at Numeric	1004_at Numeric	1005_at Numeric	1006_at Numeric	1007_s_at Numeric	1008_f_at Numeric	1009_at Numeric	100_g_at Numeric	1010_at Numeric	1011_s_at Numeric	1012_at Numeric	1013_at Numeric	10 N
5	318.4	40.2	11.2	-238.7	48.3	477.1	30.2	857.3	960.1	660.8	224.4	-7.7	246.9	6.7	21.6	- ·
6	242.4	83,9	12.5	-10.4	16.1	945.7	3.3	925.5	662.1	675.5	104.8	-23.5	180.6	6.3	2.2	
57	299.0	21.0	6.5	-161.5	23.0	920.2	18.7	366.8	911.9	1349.6	167.5	-2.8	337.0	-2.4	19.5	
18	190.0	39.5	13.8	-38.4	-2.1	738.2	22.9	452.7	645.2	1057.7	170.0	16.8	300.9	8.5	18.8	
59	245.9	88.0	50,9	-225.3	10.6	1134.2	43.8	684,4	206.8	337.3	201.8	11.6	184.7	11.1	-4, 1	
10	267.9	36.5	23.8	-47.5	44.5	538.1	49.6	592.0	239.3	955.1	205.1	-22.4	187.3	-0.3	1.7	
÷1	431.7	29.1	-1.6	-204.0	29.9	229.3	-41.4	523.7	1311.6	937.1	202.0	17.8	207.9	2.8	29.4	
42	189.1	37.9	-27.8	-180.6	21.3	199.0	41.2	279.7	314.7	755.6	226.8	46.7	295.5	-6.1	17.3	
43	238.8	41.1	37.7	-230.7	41.3	1012.9	24.1	686.7	920.5	1112.0	216.2	17.8	254.4	-3.2	13.5	
8-8	232.6	45.3	1.0	-125.5	9.6	1223.3	37.0	474.8	951.8	762.0	185.8	20.0	142.4	10.6	11.0	
45	211.6	49.4	-20.9	-162.3	7.1	1558.6	49.4	439.2	1481.3	584.1	325.2	11.7	200.8	33.2	-2.1	
46	265.3	70.1	67.7	-250.2	0.0	339.0	22.7	1120.5	986.6	658.8	264.5	6.3	169.5	0.2	2.1	
17	322.7	23.2	8.7	-115.0	67.7	255.8	88,4	1188.9	927.6	641.7	320.1	-33.8	184.1	-11.5	-0.1	
48	200.9	68.0	-1.8	-280.8	33.7	862.0	69.9	377.3	1245.3	1085.5	262.1	4.0	235.9	6.4	7.4	
19	280.1	81.4	23.2	-193.3	45.9	509.5	13.8	764.1	1697.2	491.7	269.8	25.2	233.8	18.2	-3.8	
0	244.8	45.7	32.4	-124.3	57.7	481.7	30.2	626.1	1300.8	789.2	133.3	12.0	410.2	9.8	-2.1	
1	76.9	40.2	13.0	-103.7	0.7	678.8	-10.4	860.4	1307.1	1124.5	379.6	28.2	235.2	10.0	21.3	
52	283.8	35.8	185.1	-168.5	22.5	324.7	6.5	723.0	1202.9	1260.8	259.6	-8.2	151.1	-30.3	-0.2	
53	243.2	42.5	17.3	-224.5	26.8	183.5	-23.0	662.8	1368.1	702.3	220.6	6.3	216.0	16.2	-5.6	
54	285.1	23.0	17.8	-167.9	10.9	591.7	75.6	489,8	1928.3	1195.2	275.7	15.6	204.7	5.9	-12.3	
55	198.5	87.6	10.1	-134.7	37.3	1423.6	13.0	513.9	1249.4	815.1	208.3	-8.6	277.8	12.3	16.5	
56	304.9	78.4	-8.8	-176.8	51.3	1771.7	7.4	606.8	1020.4	622.8	231.3	40.8	223.0	13.4	-0.2	
57	238.2	31.5	20.2	-230.1	32.8	733.2	-22.8	649.6	1106.9	614.3	185.2	13.0	277.9	2.8	27.2	
58	110.6	59.5	-19.8	-155.0	42.4	2766.9	48.8	520.0	867.5	788.1	194.7	-9.9	257.9	9.6	0.2	
59	293.6	45.1	3,3	-142.1	54.0	850.6	1004.6	678,9	1023.7	504.9	212.1	-14.7	187.1	-5.7	17.6	
50	160.5	35.9	-21.2	-132.7	-1.2	1513.9	-12.9	622.8	956.1	508.1	222.6	6.6	372.5	3.4	-1.7	
51	293.7	45.9	3.3	-233.0	32.1	227.9	47.5	495.7	985.5	331.0	211.5	8.7	187.0	3.7	-2.9	
52	241.6	36.8	-5.6	-67.2	29.8	634.9	-35.7	575.1	769.5	927.6	129.8	-5.8	224.1	11.4	1.5	
F3	282.7	58.8	-18.3	-200.4	-4.1	424.1	22.9	968.0	291.3	193.5	151.3	21.4	279.3	6.8	0.9	
5-4	296.6	49.5	10.4	-110.7	15.5	289.5	27.2	693.3	282.1	498.4	251.6	16.6	263.5	3.1	4.0	
55	313.4	36.0	-2.1	-202.7	26.9	328.3	31.2	601.6	260.6	730.4	214.8	30.9	154.8	0.2	7.9	
56	269.6	61.4	-16.3	-163.2	58.7	2717.9	17.8	521.3	469.9	641.1	235.6	11.5	282.2	3.8	6.0	
37	198.8	26.7	15.0	-200.6	40.0	179.5	42.3	433,7	475.4	826.0	190,1	9,3	218.6	-0.2	14.9	
58	369.7	49.3	-26.8	-187.7	18.7	935.5	11.8	475.7	347.4	969.6	164.0	9.6	124.5	11.6	6.1	
59	340.3	41.9	-17.3	-159.2	26.6	273.4	47.3	422.0	305.2	745.0	265.9	0.9	274.0	11.6	11.5	

Figure 6: Spreadsheet of the training microarray dataset

Genetic Algorithm

All the 12533 genes of 69 instances for training dataset were subjected to genetic algorithm for the purpose of reducing high dimensionality of the dataset (Feature Selection) using Weka. Weka uses its attribute selection called "Genetic Search", the mutation probability rate was set at 0.033 for all features present and crossover probability to 0.6. The population size was set at 20 individuals. These parameters are summarized in figure 6

0	weka.gui.GenericObjectEditor	×						
	ection.GeneticSearch							
About								
GeneticSearch	GeneticSearch: M							
Performs a se	earch using the simple genetic algorithm described in Goldberg (1989).							
crossoverProb	0.6							
maxGenerations	20	5						
mutationProb	0.033	5						
populationSize	20	5						
reportFrequency	20							
seed	1							
startSet								
Open	Save OK Cancel							

Figure 7: The Genetic Search Parameters

(IJARMS) 168

After setting the parameters for the genetic algorithm as shown in figure 7, Genetic Search module reduces 12533 features for both training and testing datasets to 748 features. This constitutes 94.03% reduction of the total features in the two datasets. Figure 8 shows the overall generated initial population subsets using the Genetic Algorithm. Figure 9 shows the 20 generations obtained during the process of executing the Genetic Search module and Figure 10 shows the reduced features as the final output of the Genetic Algorithm feature selection process.

Attribute selection	output .		
Run info	rmation ++++		- 1
Evaluators		wBelection.CfwBubeetEval	
Search:weka.	attributeSelect:	inn.GeneticSearch -2 20 -5 25 -C 0.4 -H 0.033 -R 20 -5 1	
Relation:	changed by hit	laing .	
Instances:	69		
Attributer:			
	ributes omitted		
Evaluation m	oderevaluate on	all training data	
Attribut	e Selection on (all input data	
Search Netho	dir.		
Gene	tic search.		
Star	intte on the r	autes	
Popul	lation size: 20		
Reb	er of generation	A#2 20	
Prob	ability of cross	auvezi 0.6	
Frob	ability of muter	Alen: 0.033	
	rt frequency: 21		
Rand	on number seed:	1	
initial popu	lation		
merit	scaled	RUBH1	
0.04524	0.00611	1 4 6 7 8 11 13 18 16 17 18 21 23 28 28 28 28 29 30 33 37 40 43 48 48 49 50 54 55 56 57 58 59 62 63 64 45 66 70 72 73 74 78 52 63 64 68 59 50 91 93 96 97	90 99 10
0.0445	0.06169	4 5 7 8 13 14 15 17 20 22 23 24 27 31 33 34 35 34 37 38 29 40 41 43 44 45 50 51 53 55 57 50 59 40 44 45 64 65 97 07 72 73 74 75 77 76 79 51 53 54 55 59	90 91 9
0.06334	0.05474	2 4 9 10 10 14 15 19 20 21 27 30 33 34 39 41 45 46 47 48 51 53 54 55 57 58 59 60 65 71 72 52 54 55 91 52 59 96 99 102 100 104 106 107 106 109 111 114 1	116 117 1
0.06004	0.03499	9 28 33 37 100 104 123 136 211 253 272 288 292 298 304 321 326 387 388 394 400 424 433 460 475 488 531 575 613 630 634 704 708 712 747 838 850 937 987	1075 111
0.0667	0.07488	19 20 21 27 31 52 36 39 51 53 45 48 72 73 80 86 67 91 99 100 101 104 107 110 113 114 124 125 133 134 135 137 138 142 145 148 148 150 152 157 147 170 171 17	
0.06324	0.05416	4 9 16 18 21 23 45 51 53 67 68 70 77 85 204 105 107 111 112 122 137 149 151 153 155 158 163 168 176 165 182 180 196 206 207 209 211 212 223 229 234 240	
0.063	0.0527	3 5 6 7 10 11 12 14 18 20 22 23 24 26 28 30 34 42 43 44 48 47 50 51 59 66 49 72 74 76 77 78 79 82 84 90 91 93 94 95 96 97 90 101 102 103 106 108 114 11	15 116 11
0.04145	0.0434	11 12 14 19 29 40 45 51 54 59 64 69 72 75 78 81 90 90 94 110 122 124 129 138 139 143 151 154 166 169 170 172 181 187 190 197 200 202 206 207 208 215 22	
0.04825	0.09076	3 6 10 11 15 18 23 25 35 38 40 46 47 55 42 48 73 80 83 88 89 48 103 106 109 112 118 120 126 131 133 134 139 142 145 146 146 149 178 181 183 186 192 213 214	
0.06435	0.06078	2 2 4 5 7 10 11 12 14 15 17 18 19 20 20 24 28 29 30 32 33 35 37 39 41 42 44 45 51 52 59 41 42 45 67 70 71 72 72 12 83 68 67 19 11 12 45 67 79 11 12	/ 105 108
< 1			

Figure 8: Generated initial population subsets

Generation:	20	
merit	scaled	subset
0.08001	0.14424	14 19 28 37 39 47 98 100 113 117 123 198 220 223 225 229 247 298 301 321 329 380 415 440 465 502 517 532 536 559 567 592 619 687 690 691 720 749 751 752 761
0.08001	0.14424	14 19 28 37 39 47 98 100 113 117 123 198 220 223 225 229 247 298 301 321 329 380 415 440 465 502 517 532 536 559 567 592 619 687 690 691 720 749 751 752 761
0.07331	0.083	1 6 9 14 19 27 28 37 39 47 50 53 55 69 100 102 109 113 115 117 123 129 131 137 141 165 169 198 200 210 220 223 225 226 229 247 250 251 259 261 263 270 279 28
0.07339	0.08369	3 5 10 12 13 14 19 20 21 26 28 30 31 34 36 37 47 49 51 53 54 55 56 60 62 67 71 74 77 93 98 100 104 113 117 118 120 121 123 126 129 133 136 138 144 145 152 15
0.07207	0.07165	11 14 19 28 37 39 47 98 100 113 117 123 155 198 200 205 215 219 220 223 225 229 230 247 255 264 298 301 313 321 329 363 376 380 392 396 397 412 415 416 440
0.07454	0.09424	2 10 13 17 18 28 31 32 37 38 39 40 45 46 47 49 51 56 58 68 74 76 79 81 84 91 93 97 98 100 109 112 113 115 116 117 119 121 123 124 126 128 131 133 141 142 15
0.07126	0.06425	3 5 12 13 14 19 20 21 26 28 30 31 34 37 47 49 54 56 60 62 67 69 71 74 77 93 98 100 113 117 118 120 121 123 129 133 138 144 145 152 156 157 161 163 168 171 1
0.06829	0.03711	9 14 19 22 28 31 35 36 37 39 51 54 71 81 83 92 98 100 113 114 117 123 149 153 155 159 160 167 171 173 196 198 204 210 217 218 220 223 225 226 227 229 239 24
0.06885	0.04219	3 12 14 19 28 33 39 40 47 48 65 98 100 101 113 117 118 144 175 184 198 209 215 216 220 221 223 225 229 247 253 256 260 272 281 293 298 305 321 322 340 344 3
0.07575	0.10532	14 19 36 37 47 58 65 69 71 65 91 98 100 104 113 117 123 140 141 149 159 162 166 167 169 188 198 205 214 220 223 225 229 238 247 270 275 290 298 301 309 321
0.0715	0.06642	2 3 11 13 14 15 17 18 28 31 37 38 39 40 46 47 49 51 56 58 60 65 76 79 81 83 84 91 93 98 100 113 117 119 120 123 124 126 128 131 133 141 143 149 150 161 166
0.07163	0.06761	3 12 14 19 28 39 47 65 69 98 100 101 113 117 118 149 168 173 175 184 194 198 205 209 211 216 220 221 223 225 229 247 250 253 256 260 281 290 293 298 301 305
0.06554	0.01188	14 21 28 37 39 47 51 52 58 66 69 91 92 99 100 107 113 114 115 117 118 120 123 127 128 130 139 155 166 171 173 183 195 198 199 219 220 223 225 226 234 240 24
0.06758	0.03052	2 8 9 13 15 17 18 28 31 37 38 39 40 42 45 46 47 49 51 56 58 60 62 74 76 79 81 82 84 91 93 98 100 113 117 119 123 124 126 131 133 141 142 150 152 161 163 166
0.07663	0.11337	4 9 14 17 19 28 37 39 43 45 58 60 61 69 76 82 84 98 100 113 123 136 171 192 198 199 211 220 223 225 229 232 237 247 278 298 301 302 311 316 321 329 334 335
0.06937	0.04696	1 14 15 19 24 27 28 37 39 47 53 55 69 74 80 90 98 100 102 113 117 123 129 131 137 145 148 165 169 177 193 198 216 223 225 226 229 236 240 247 251 252 257 27
0.07513	0.09963	1 8 10 14 15 16 17 19 25 28 36 39 43 45 47 60 87 95 98 100 113 118 120 123 131 137 149 158 167 185 194 198 203 208 218 220 221 222 223 225 229 234 247 250 2
0.06775	0.03209	2 3 10 13 17 18 27 28 31 32 37 38 39 40 46 47 49 51 56 58 68 74 76 79 81 84 85 91 93 98 105 109 113 115 119 121 123 124 126 131 133 141 142 150 154 161 166
0.07087	0.06069	4 14 19 36 37 39 47 58 65 66 68 70 76 88 91 98 100 113 117 123 149 162 167 169 187 198 220 223 225 229 238 247 261 270 275 265 290 298 301 309 321 329 353 3
0.06897	0.04333	14 21 26 28 37 39 42 47 51 55 60 65 69 71 90 91 92 98 100 113 114 115 117 118 120 123 126 127 130 138 139 171 183 195 198 199 214 219 220 223 225 226 231 241

Figure 9: Subset of 20 Generations

Attribute selection output	
Attribute Subset	<pre>Cvaluator (supervised, Class (nominal): 12534 Class):</pre>
CFS Subse	t Evaluator
Including	locally predictive attributes
Selected attribut	Her: 14,19,28,37,39,47,96,100,113,117,123,196,220,223,225,229,247,296,301,321,329,380,415,440,445,502,517,532,536,559,567,592,419,687,690,691,720,749,751,752,761,762,789,840,401,1012_at
	1017_at
	1025_g as
	1003 g.at
	1035_g_at
	1042 at
	log j. at
	1080 f at
	1102 s at
	1106 s at
	lill at
	117 at
	1200_at
	1203_at
	1205_at
	1209_at
	1227_g_at
	1275_at
	1270_at
	1297_at
	1506_at
	1353_g_st
	1385_at
	1400_at
	1431_at
	1465 <u>s</u> at
	1479_g_at
	1492 <u>f</u> at
	1496_at
	1518 at
	1525_9_at
<	>

Figure 10: Final reduced features

Multilayer Perceptron

The performance of a good classifer become more pronouced when subjected to a number of salient features. The reduced features presented in the final output of the Genetic algorithm were supplied as input to the multilayer perceptron accessed through the Classify feature of Weka. MLP model was trained with 69 instances consisting of 748 reduced attributes.Figure 11 shows the weight adjustement during the training of the multilayer perceptron neural network classifer. Figure 12 shows the neural network classified instances, Kappa statistic, Mean absolute error, Root mean squared error, Root relative squared error and total number of instances used for the training. The detailed accuracy by class section in figure shows that the classifer received good training with 100% accuracy and it is much ready for classification with testing dataset.

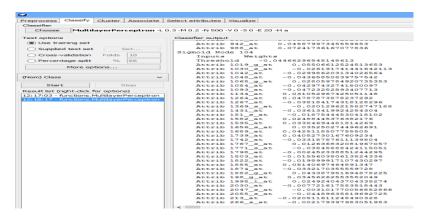


Figure11 : Adjustment of weight during training of the model

✓	Weka Explorer	
Preprocess Classify Cluster Associa	te Select attributes Visualize	
Classifier		
Choose MultilayerPerceptron	a +L 0.9 -M 0.2 -N 500 -Y 0 -S 0 -E 20 -H a	
Test options	Classifier output	
 Use training set 	Attrib 945_at -0.006223219591550792	
 Supplied test set Set 	Attrib 963_at -0.03257928735093213 Attrib 967_g at -0.012220201557473556	
O Cross-validation Folds 10	Class Mesotheliona	
O Percentage split % 66	Input	
More options	Node 0 Class ADCA	
	Input	
(Nom) Class	V Node 1	
Start Stop		
Result list (right-click for options)	Time taken to build model: 265.03 seconds	
13:12:20 - functions.MultilayerPerceptro		
	=== Evaluation on training set ===	
	Stimity	
	Correctly Classified Instances 69 100 %	
	Incorrectly Classified Instances 0 0 %	
	Kappa statistic 1 Mean absolute error 0.0014	
	Rean absolute error 0,0014 Root mean squared error 0,0021	
	Relative absolute error 0.5462 %	
	Root relative squared error 0.5521 %	
	Total Number of Instances 69	
	Detailed Accuracy By Class	
	TP Rate FP Rate Precision Recall F-Measure ROC Area Class	
	1 0 1 1 1 1 Mesotheli	oma
	1 0 1 1 1 1 ADCA	
	Weighted Avg. 1 0 1 1 1 1	
	Confusion Matrix	
	a b < classified as	
	10 0 a = Mesothelioma 0 59 b = ADCA	
	0 59 1 B = ADCA	_

Figure 12: Trained neural network classifier

Figure 12 shows the accuracy of the neural network classifier when subjected to testing dataset. The classifier acheived accuracy of 97.5% with false positive (FP) rate 1.3%. The confusion marix is shown in Figure 14 .Seventy eight (78) instances were correctly classified out of the 80 instances used for testing. The model also achieved good true positive (TP) rate, FP rate, precision, recall, f-measure and receiver operating characteristic (ROC) area.

<u>۵</u>	Weka Explorer 💛 🗕 🗆 🗙
Preprocess Classify Cluster Associate Select attributes Visualize Classifier Choose MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -5 0	E 20 +H a
	Casefer output Time taken to build model: 5.19 seconds
Status OK	Log 💉 x t

Figure 13: Performance of neural network classifier

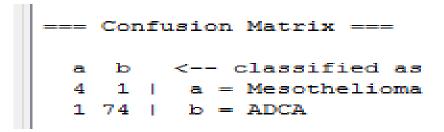


Figure 14: Confusion matrix

Conclusion and Recommendations

A major genetic algorithm parameter change would be to increase the population size. By increasing the population size, the algorithm would perform a more thorough search of the solution space and would be more likely to locate the global minima. The results presented in this research revealed that the proposed hybrid GA/MLP performs better with over 90% accuracy when used to classify microarray dataset of lung cancer. Detection of cancer disease is an important issue in the research community. A lot of efforts have been committed to develop models that are capable of diagnosing cancer. However, these models have one issue or the other that are needed to be addressed. In this dissertation, a hybrid model that combines the optimization power of Genetic algorithm for reduction of high dimensional microarray data and MLP for classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung was proposed. The solution found by the combine Genetic Algorithm (GA) and Multilayer Perceptron (MLP) algorithm performed effectively well. The GA only focuses on the reduction of high dimensionality of microarray dataset and NN focuses on the accurate classification. The mutation probability rate of the GA was set to 0.033 for all features present and crossover probability to 0.6. The population size was set at 20 individuals. The GA reduced the 12,533 attributes in the microarray dataset to 748 attributes. The reduced microarray dataset was used to train the multilayer perceptron NN classifier. The trained classifier achieved 97.5% accuracy when evaluated with the testing microarray dataset. To further improve the proposed model presented in this paper for hybridization, the following recommendations may prove useful.

- 1. Different parameters can be used to configure the Genetic algorithm for the purpose of improving its feature selection process.
- 2. The proposed model can be further validated using different microarray cancer datasets in order to ascertain its efficiency.
- 3. For quicker feature selection, filter and wrapper methods in Weka can be employed to perform feature selection due to the complexity and robustness of the Genetic algorithm.
- 4. Apart from the feed-forward back propagation algorithm used in this work for training, other algorithms can also be employed to train the multilayer perceptron, neural network classifier.

References

Abdul Jaleel, J., Sibi, S., & Aswin, R. B. (2012). Artificial Neural Network Based Detection ofSkinCancer.International Journal of Advanced Research in Electrical, Electronics andInstrumentationEngineering, Vol. 1 (3).Cancer.

Abraham, K. (2012). Machine Learning Algorithms for Cancer Diagnosis.

Alireza, O., & Bita, S. (2011). A Computer Aided Diagnosis System for Breast Cancer. IJCSI International Journal of Computer Science Issues, Vol. 8 (2).

American Canser Society. (2020). Cancer Facts & Figures 2014. Cancer Research UK. (2020). The Lung.

Christos, S., & Dimitrios, S. (2003). NEURAL NETWORKS.

Ganesan, N., Venkatesh, K., Rama, M. A., & Malathi, P. A. (2010). Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data. International Journal of Computer Applications, Volume 1 - No. 26

- Khalid, R., & Atif, N. H. (2013). A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction Based on Microaray Data. International Journal of Bioinformatics Research and Applications .
- Kyu-Baek, H., Dong-Yeon, C., Sang-Wook, P., Sung-Dong, K., & Byoung-Tak, Z. (2002). Applying Machine Learning Techniques to Analysis of Gene Expression Data : Cancer Diagnosis.

Medical News Today. (2019). What Is Cancer? What Causes Cancer?

Jimoh.M.K (2015). Hybridization Of Genetic-Neural Algorithms For Cancer Disease Diagnosis Using Microarray Data. University of Ilorin, Ilorin.

National Cancer Institute. (2020). What Is Cancer.

Shelly, G., Dharminder, K., & Anand, S. (2019). Data Mining Classification Techniques Applied for Breast Cancer
 Diagnosis and Prognosis. Indian Journal of Computer Science and
 Engineering (IJCSE), Vol. 2 (No. 2).

Tom A., v. (2020). A History of DNA Microarrays. Pharmaceutical Discovery, Vol. 5(1), 23-46.

Vitoantonio, B., Giuseppe, M., Filippo, M., Angelo, P., & Stefania, T. (2006). Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis: a Distributed Approach. Engineering Letters, 13 (3), 67-81.

World Health Organization. (2013). Media centre. WHO.

World Health Organization; International Agency for Research on Cancer;. (2014).