CROSSBREEDING OF NEURAL AND GENETIC ALGORITHM FOR CANCER DISEASE DIAGNOSIS USING MICRO-ARRAY DATA

By

AHMED ABIODUN TAOFIK (T. A. Ahmed)
Department of Computer Science Kwara CAILS, Ilorin

ABSTRACT

This paper attempt a shift in paradigm from conventional methods by formulating cross breeding model of genetic algorithm and neural network for optimization of relevant features of the genes and classification of cancer disease respectively. Microarray data is to be considered as a dataset.

The paper therefore, adopt crossbreeding model of Genetic Algorithm and Neural Network, it was developed and simulated in Weka environment using microarray cancer dataset. The solution found by the combined Genetic Algorithm and Neural Network performed effectively well.

The results presented in this paper revealed that the proposed crossbreeding of Genetic Algorithm and Neural Network performs better with over 97% accuracy when used to classify microarray dataset of lung cancer.

Keywords: Cancer, Diagnosis, Genetic Algorithm, Neural Network, Hybridization or Cross breading

Introduction

Detection of cancer disease in its early stage 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., 2017). 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 (2017), 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.,2017) as quoted in Khalid andAtif (2017). 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 M.K.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 reveal by American Cancer Society (2017) 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 (2017).

Problem of the Statement

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 (2017), 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 2017. The most common causes of cancer death are cancers of lung (about 3 million deaths), liver (about 1million 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, 2017). Projections based on the GLOBOCAN 2017

Aim and Objectives

The aim of this research is to formulate and implement a crossbreeding model that combines the power of Neural Network and Genetic Algorithm for cancer disease diagnosis based on microarray data.

The objectives are:

i. to formulate genetic-neural crossbreeding model for 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

Scope of the Study

The scope of this research is to apply neural networks 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 (2017), 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 (2017), 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.

History of Genetic Algorithms

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.

History of Artificial Neural Network

Neural networks 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 2017)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.

History of Microarray

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 A., 2005). The first arrays, created in the mid 80s, 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.

Review Of Related Works

Khalid and Atif (2017) 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

Methodology

Proposed System Framework

The proposed system consists of different modules and divided into two stages as shown in Figure 3.1 and 3.2. 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

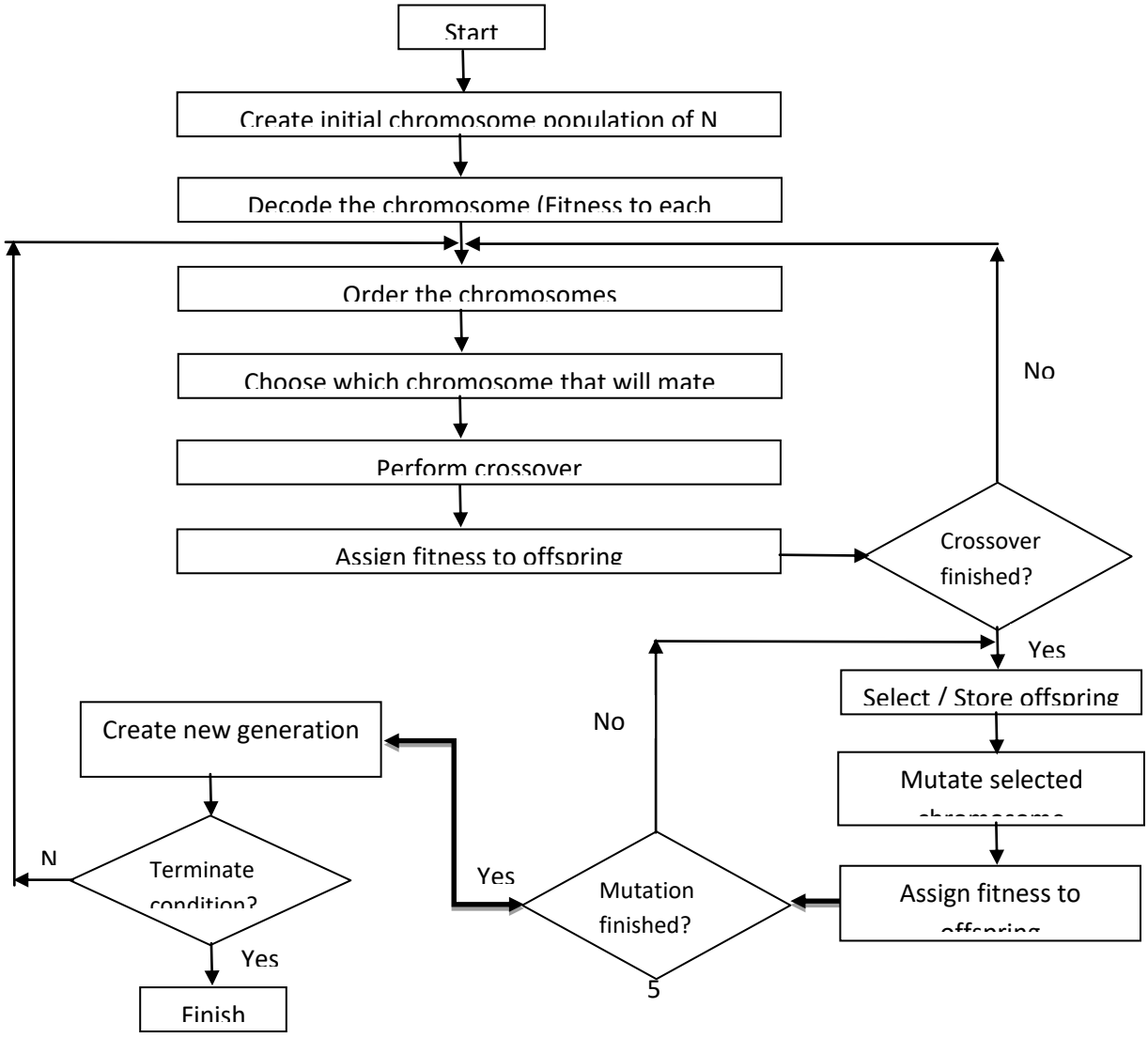

Figure 3.1: Proposed system framework (Genetic Algorithm)Stage 2: Neural Network for the classification of cancer disease

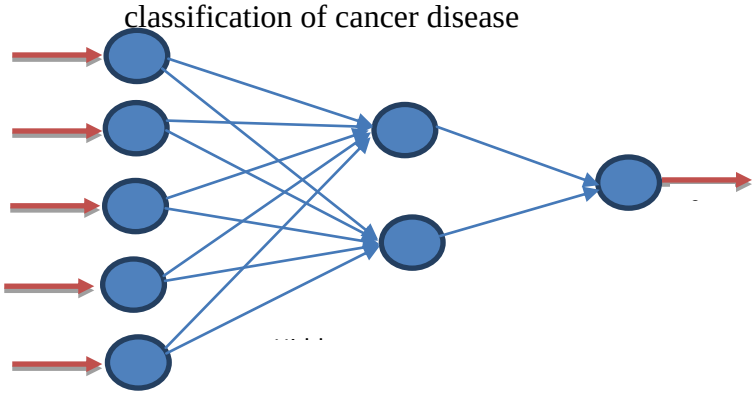

Figure 3.1: Proposed system framework(Neural Network)

The above process is repeated until some condition is satisfied (Rahul, Narinder, and Yaduvir, (2017)). 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

Discussion and Findings

The goal of this research is to crossbreed genetic algorithm and neural network 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 neural network 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.

System Specification

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 Microarray Dataset

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 4.1 (a and b) and 4.2(a and b) respectively. Each sample is described by 12533 genes and figure 4.3 shows the visualization of the microarray dataset. Figure 4.4(a and b) show the spreadsheet copy of both training and testing dataset respectively.

Figure 4.1(a): The training set contains 69 of tissue samples, 10 MPM and 59 ADCA

Figure 4.1(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

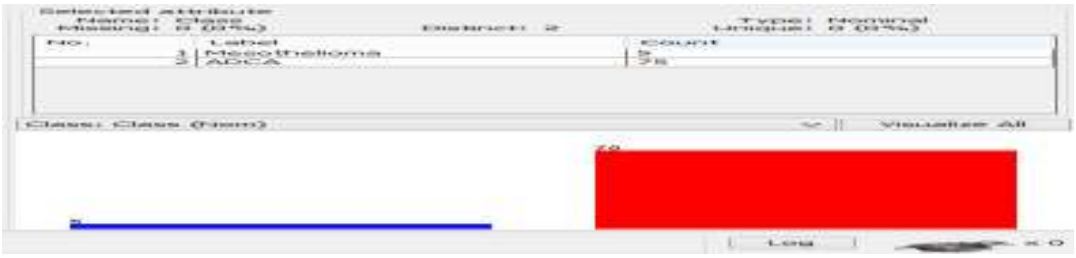

Figure 4.2(a): The testing set contains 80 of tissue samples,5 MPM and 75 ADCA

Figure 4.2(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

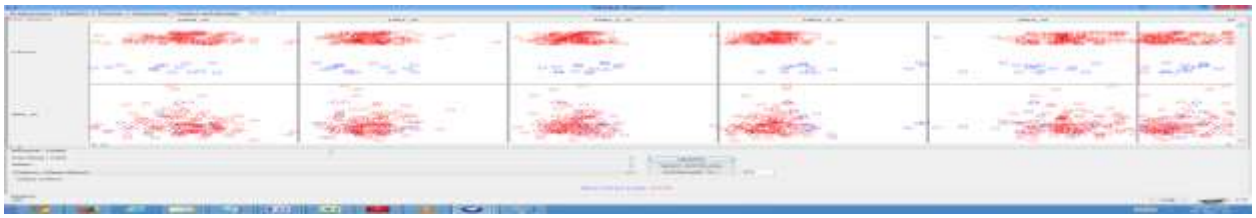

Figure 4.3: The visualization of the microarray dataset

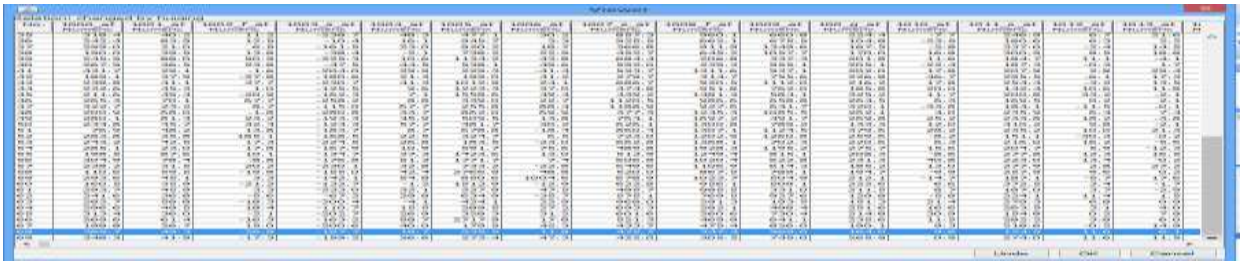

Figure 4.4(a): Spreadsheet of the training microarray dataset

Result of 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 4.4.

Figure 4.5: The Genetic Search Parameters

After setting the parameters for the genetic algorithm as shown in figure 4.5,GeneticSearch 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 4.6 (a) shows the overall generated initial population subsets using the Genetic Algorithm. Figure 4.6 (b) shows the 20 generations obtained during the process of executing the GeneticSearch module and Figure 4.6 (c) shows the reduced features as the final output of the Genetic Algorithm feature selection process.

Figure 4.6 (a): Generated initial population subsets

Figure 4.6 (b): Subset of 20 Generations

Figure 4.6 (c): Final reduced features

Result of Neural Network

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. Neural network model was trained with 69 instances consisting of 748 reduced attributes.Figure 4.7 shows the weight adjustement during the training of the multilayer perceptron neural network classifer. Figure 4.8 shows the neural network classifer during training with 100% correctly classified instances. The figure also shows the number of incorrectly 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 4.8 shows that the classifer received good training with 100% accuracy and it is much ready for classification with testing dataset.

Figure 4.7: Adjustment of weight during training of the model

Figure 4.8: Trained neural network classifier

Figure 4.9 shows the accuracy of the neural network classifier when subjected to testing dataset. The classifer acheived accuracy of 97.5% with false positive (FP) rate 1.3%. The confusion marix is shown in Figure 4.10.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.

Figure 4.9: Performance of neural network classifier

```
=== Confusion Matrix ===

  a  b   <-- classified as
  4  1 |  a = Mesothelioma
  1 74 |  b = ADCA
```

Figure 4.10: Confusion matrix

Summary, Conclusion and Recommendation

Summary

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, an hybrid model that combines the optimization power of Genetic algorithm for reduction of high dimensional microarray data and Neural network 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 Neural Network (NN) 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.

Conclusion

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 dissertation revealed that the proposed hybrid GA/NN performs better with over 93% accuracy when used to classify microarray dataset of lung cancer.

Recommendations

To further improve the proposed model presented in this paper, 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. 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.
3. 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.
4. The proposed model can be further validated using different microarray cancer datasets in order to ascertain its efficiency.

REFERENCES

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

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. (2017). *Cancer Facts & Figures 2014.*

Cancer Research UK. (2017). *The Lung.*

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

Eric, R. S. (2017). *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. (2017). *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. (2017). *What Is Cancer.*

Shelly, G., Dharminder, K., & Anand, S. (2017). 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. (2017). A History of DNA Microarrays. *Pharmaceutical Discovery, Vol. 5* (7).

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).

World Health Organization. (2014). *Media centre.* WHO.

World Health Organization; International Agency for Research on Cancer;. (2013). *Latest world cancer statistics.* WHO.