Enhancing the Efficiency of Parallel Genetic Algorithms for Medical Image Processing with Hadoop

D. Peter Augustine
Associate Professor
Christ University
Bangalore 560029

ABSTRACT

In this paper, there is in-depth analysis of the parallel genetic algorithms used for segmentation of brain images and how their efficiency varies in the cloud setup with Hadoop. Since the current health care industry is moving towards the utmost usage of cloud to make the data available round the clock for the analysis, it is mandatory that the efficiency of the analysis also to be enhanced to produce the accurate result. Here, the focus is on the study of medical image processing that too narrowed down to the brain images with the help of parallel genetic algorithms in the cloud environment. The study aims to help the researchers to augment the competence of the algorithm when it functions in the remote cloud setup.

General Terms

Medical Image Processing, Parallel genetic algorithms, Health care, Health Care Applications, Hadoop, Cloud Computing.

Keywords

Efficiency of parallel genetic algorithms, Image processing in cloud environment, Brain Image processing, Hdoop's Map Reduce.

1. INTRODUCTION

Imaging is a powerful mechanism to visualize an area of interest to analyze and get the desired attributes from them. It is same in the field of medical imaging. Medical imaging is a powerful tool in the field of medicine to take a further step to diagnose and cure diseases. The imaging techniques like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultra Sound imaging (US) and others are very effectively providing the information about the anatomy of the human body. These technologies are very supportive in diagnosis and treatment planning. There are various computer algorithms to describe the anatomical structures and other regions of interest. Researchers are working on improving the performance of these algorithms by incorporating different models. These algorithms called as image segmentation algorithms, contribute a maximum in medical imaging applications. With the support of next generation technologies like Hadoop and Big Data Analytics, one can find the way of improving the efficiency of algorithms in medical image processing.

In the section 2, the analysis of medical image processing is done where one can find the ideas of the same. Parallel Genetic Algorithm (PGA) is explained with its advantages while equipping it in the field of image processing in the section 3. Section 4 explains Big Data Analytics in the image processing which paves the way for PGA to be incorporated in Hadoop. Study of PGA implementation is in the following section 5. Section 6 describes the Future Research Directions which can give ideas for the researchers in the same field of research. Section 7 concludes the paper.

2. MEDICAL IMAGE PROCESSING

Medical image processing (MIP) is a blessing for the human race in the field of computer science to detect the problems in the human anatomy. In the present fast growing and competing technological word, medical imaging tries to make diagnosis as a second perfect doctor. In automated medical diagnostic systems, MRI (magnetic resonance imaging) gives better results than computed tomography (CT) as MRI provides greater contrast between different soft tissues of human body.

Many disorders of brain exist these days out of which degenerated tissue (tumor) is spreading at an alarming rate. Even more than 400000 persons per year in the world (based on the World Health Organization (WHO) estimates) are suffering from this disorder.

The sequence of operations for detection of tumor area in MRI images of brain contain various steps like preprocessing, post-processing, classification and volume calculation and these are shown in following steps.[14]

Step 1. Image Collection

Step 2. Preprocessing

Step 3. Image Enhancement

Step 4. Post-processing

Step 5. Segmentation using threshold value

Step 6. Tissue classification.

2.1 Image Preprocessing and Enhancement

Preprocessing is the mechanism to improve the reliability of optical inspection of the desired area in the image. For instance, to detect the tumor in the brain image, the MRI images of brain are acquired first and then they must undergo pre-processing in order to extract the necessary information. The image quality may be in uncertainty due to various artifacts like out of focus, presence of noise, distortion of optical systems, the relative motion between the camera and the scene etc. There are different preprocessing techniques to get the image out of theses artifacts.

The main focus of image enhancement is to make the interpretability of information scattered in images and in turn it can provide better input for other automated image processing techniques. The film artifacts like labels and marks on the MRI image and the high frequency components are removed during enhancement stage. MRI image is converted into standard image without noise, film artifacts and labels after image enhancement. There are different image enhancement techniques like median filter, Gaussian filter, normalization methods to do get the standard images.

2.2 Image Segmentation

The Segmentation of an image involves the splitting or separation of the image into sections of related characteristics. This is an important mechanism which extracts various characteristics of an image. These characteristics can be merged or divided in order to form an object of interest for further analysis and interpretations.[9]

Image segmentation includes clustering, thresholding etc. The most recurrent technique used for segmentation is Global Thresholding method which is simple and effective. Threshold segmentation is technique of converting gray scale image into binary image.

There are various techniques used by image segmentation like edge and region detection, statistical classification, thresholding or by combining any of them and the outcome will be a set of classified elements. Region-based or edge-based techniques are commonly used in segmentation for the best results.[5]

2.3 Feature Extraction

From the phrase itself it explicitly tells that it is a technique of pulling out meaningful features from input image. For instance, in the study of brain tumor detection, degenerated tissue is to be extracted from the segmented image with the presence of minimal surplus elements and it represents reduction of dimensionality. It is not an easy task to extract useful information since it is mandatory for volume calculation.

2.4 Tissue Classification

This is another technique which classifies the tissues into two classes namely Normal and Abnormal tissue. It begins with more distinct features and steadily adding features of less distinct, until classification performance cannot proceed further. SVM, Artificial neural network, k-Nearest Neighbor (k-NN) etc are some of the classification techniques.

3. PARALLEL GENETIC ALGORITHM FOR MEDICAL IMAGE PROCESSING

Segmentation models proposed in various studies depend on the particular application to visualize images, modality like CT, MRI, etc. and different factors. There are various influencing aspects differ from one part of anatomy to another part of anatomy. The algorithms work very well for one particular segmentation process may not yield the accurate results for another. The influencing factors like motion, partial volume effects, noise and other environments may interfere in the results of the algorithms. An algorithm which can deal with one of these artifacts may not go well with other factors. This is one of the major challenges in the image segmentation. Because of these shortcomings of segmentation algorithms, selecting one particular algorithm for segmentation is a difficult but a necessary task.

Genetic algorithms are outcome of Darwin's theory about evolution. As it has been described earlier this model gives solution to various problems.[16]

The set of solutions represented by chromosomes called population are the input to the algorithm. The extracted solutions from one population are taken then used to form a next generation children which is considered as new population. The new population referred as offspring is considered as the better than old population. The offspring will be chosen based on their fitness value. The above process is repeated until the desired outcome based on the condition which was referred as a key value later.[4]

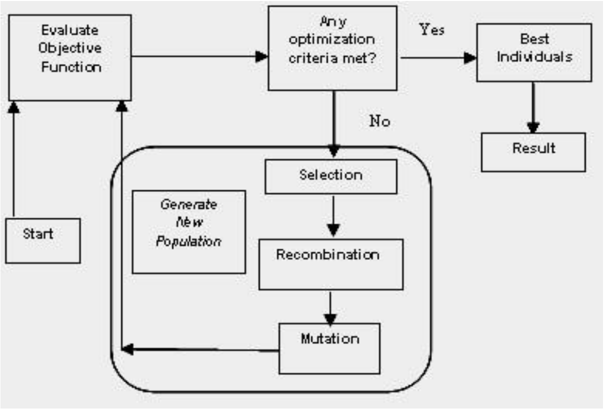

Fig 1: GA Process Flow

In the case of medical image segmentation, Genetic Algorithms (GAs) functioning based on the standards of natural selection and genetics exhibit efficiency in search methods. GAs have already produced acceptable solutions in the fields of business, engineering and science. Even though GAs are providing expected solutions in different areas, when it comes to the bigger and difficult problems they couldn’t function efficiently in terms of execution time. Taking the benefits of GAs into consideration, researcher can add little modification by making them parallel, GAs can show drastic decrease with respect to execution time.[15]

In general, Parallel Genetic Algorithms (PGAs) are easy to implement and gives better performance. As it is explained in the above sections, the data space of medical image is huge and vast for processing. The ultimate challenge with new diseases and their diagnosis is a tedious task for the doctors. But there are voluminous image data which can support the analysis to derive a conclusion. The parallel mechanism is mandatory for analyzing these huge data to get the outcome faster. By looking into the benefits of PGAs, it is time for us to effectively incorporate PGAs for MIP.[9]

Usually the parallel processing works by dividing the given chunk of image data into smaller chunks as sub populations and giving them for execution. Sometimes even the larger chunks of data are also considered individually and run in different machines parallel.

There has been considerable success in active model-based segmentation and which has frequently been used in medical image processing. Even though active model-based segmentation is effective in speed, still its performance can be increased by implementing parallelization.

The types of PGAs and their advantages are described in the following sections.

3.1 Classification of PGAs

3.1.1 Global Genetic Algorithm

It is a common and simple evolutionary algorithm in which the entire population is considered as a sole indivisible offspring generation from where the new cross-over generation takes place. The chromosomes are stored in the shared memory and algorithm executes the chromosomes on a shared memory multiprocessor. The processor only extends definite chromosomes.[12]

In the second model of global genetic algorithm, execution is performed on a shared memory computer where master-slave relations enforced. Since a master handles all the slaves there is a possibility of slave to be idle while waiting for the master to finish process.

3.1.2 Migration Genetic Algorithm

It is considered as roughly distributed parallel genetic algorithm. The entire population is divided into several sub-populations and each is considered as a reproduction space which will generate offspring. In order to get a good offspring based on the expected attributes, sometimes chromosomes are allowed occasionally to move from one sub-population into another. Individual chromosomes can be transferred in only one direction, only between the neighboring sub-populations.

3.1.3 Dispersed Genetic Algorithm

It is termed as fine distributed Parallel Genetic Algorithm. The entire population is considered as one continuous structure in this type. To enforce this, a huge number of parallel computers are interconnected to each other in a two-dimensional topography is used in such algorithms.

The pseudo code for the fine-grained parallel genetic algorithm is given in the following:**Error! Reference source not found.**

```
for each node do in parallel
    generate an individual randomly
end parallel do
while not stop_criterion_satisfied do
    for each node do in parallel
            evaluate the fitness of the individual
            get the fitness values of four neighboring individuals
            find out the optimum fitness value
            get the neighboring individual
            corresponding to optimum fitness
            uniform crossover with the local
            individual according to the crossover rate
            mutate the individual according to the mutation rate
    end parallel do
    test the stopping criteria
end while
```

3.2 Seven Reasons to use Parallel Genetic Algorithms

1. Works on a coding of the problem (least restrictive)
2. Even if there is no hardware support for parallelization, PGAs yield better search.
3. They are robust since they are independent of the problem by nature.
4. Alternative solutions can be reached to the problem.
5. Whether the data sets are residing at remote or local, parallelization is possible without doubt.
6. They support parallel search from numerous points at a time in the data space.
7. It is proven in many cases that PGAs are more efficient than SGAs.

4. OVERVIEW OF BIG DATA ANALYTICS FOR MEDICAL IMAGE PROCESSING

Before getting into Hadoop's benefits for PGAs, let us look into the creation of Medical Images every second in the real world. One of the major challenges in the big data platforms is data acquirement. One should be sure that the latency is low in capturing the data and using simple query to process larger volume of data since these systems handle large degree of data.[13]

The accumulating data are of larger in size as the images are stored continuously in the real world. Because of this reason, it is mandatory for the system to take and process the data from the original storage location. Processing these larger volumes of data and meanwhile keeping the processed data on the original data clusters can be facilitated by Apache Hadoop which provides the platform for the same.

The present visualization devices and methods produce 3D images of anatomical structures exceptionally precise and high quality. But it is a very big question about their usage for the efficient investigation since the images pop up with a great number of anatomical organs. Especially in the research on image analysis in the field of medicine is nerve breaking since it deals with human life. Image segmentation is a technique to extract the region of interest in the anatomical structure to get a clear idea about the image.[8]

It is a well known fact that Image segmentation is a burdensome process because of very large size of image datasets, complexity of the image with respect to the anatomical structure, and variation of the anatomic organs. Sometimes because of artifacts like noise and low contrast in the images may be the hindrances for the boundaries of anatomical structures to be inaccurate and disconnected. So the innermost hidden problem in segmentation is to extract the boundaries of organs or region of interests and to separate it out from the remained datasets. The following points will helps to recognize how of Big Data Analytics (BDA) can help us in the area of Medical Image Processing.

Data Organization: Since the data's are of larger volume of size, the system needs to take and process the data from the original storage location. Apache Hadoop provides a technology to process these larger volumes of data and at the same time keeping the data on the original data clusters.

Data Processing: In big data processing the data must be process in a distributed environment. The requirement for analyzing data such as medical images requires statistical and mining approach for analyzing the data. The processed images with the expected output in a fast as well as efficient manner take higher priority.

There are few basic requirements for an expected outcome from Medical Image Processing while with Hadoop.

1. The storage with consistency and scalability.
2. Infrastructure for processing must be Reliable and scalable
3. Efficient algorithms to retrieve required information must be highly available.

4. Storage to acquire real time statistics must be Scalable and readily available.

Since Big Data Analytics (BDA) can not only observe the responses on process, but it can check whether any particular groups react differently which enables us to check the chromosomes of different characteristics from different patterns. BDA has numerous better advantages than traditional clinical analysis approach since it does analysis in real time with variable population. BDA can support dynamic sample size and attributes with the rising image data. So BDA can generate faster and better up-to-date results.

5. EFFICIENCY ENHANCEMENT FOR PARALLEL GENETIC ALGORITHMS WITH HADOOP

[3]There are varieties of challenges faced by the healthcare profession in processing the data to deliver the prolific service to everyone involved in it. Processing of medical images is one of the areas among them. The recent technology Hadoop provides solution to process the image data which can be called as Big Data due to its voluminous amount in the parallel and distributed environment. Hadoop can enhance the speed of the parallel genetic algorithm in turn it can produce better results of analysis.

The paper[1] "A Parallel Genetic Algorithm Based on Hadoop MapReduce for the Automatic Generation of JUnit Test Suites" by the authors Linda Di Geronimo, Filomena Ferrucci, Alfonso Murolo and Federica Sarro is taken as a base paper for the study and proposal of the following setup for enhancing parallel genetic algorithms for Medical Image Processing.[8]

5.1 Hadoop and Map Reduce

Hadoop gives the great push for distributed parallel processing on large degree of data across servers varying form inexpensive to high level. These servers can be scaled beyond limits and used for storing and processing the data by Hadoop. There is no data to be considered for Hadoop as excessively large. There is a fast accumulation of image data every second in the real medical world for the analysis and diagnosis. Hadoop can make it easier for the analysis which was considered as a great challenge because of its efficiency and effectiveness.[6]

With the distributed file system, Hadoop is highly capable of storing files on different systems throughout the cluster. Hadoop's competence to hide the files' residing places in the cluster can make the user to feel as if they work in the local system.

Map and Reduce tasks of Hadoop are run at the systems where the data to be processed is residing. It avoids copying of data between systems. The system works well where there are large clusters than small clusters. The MapReduce framework works well competently when the data for the execution resides in the systems where process takes place.

5.2 PGA and Hadoop for Medical Image Processing

It is a well established fact that the parallel processing always improves the performance in terms of maximizing the searching space effectively and the reducing the computation time of the data. As per the advantages of PGAs discussed in section 3, one can incorporate them into the Hadoop Map Reduce to make them effective in processing images. The following algorithm outlines the sequence of process done with the PGA in the suggestive model based on the earlier researches.[10]

1. Storage of medical images in HDFS
2. Retrieving images from HDFS and they are sent to parallel genetic algorithm (PGA)
3. PGA works on it and makes the sets for parallelization to input to be given to map reduce.
4. Splitting the input population from PGA.
5. Distributing splits to Mapper.
6. Mapping (n mappers will do mapping).
7. Reducing and storing into HDFS.
8. If the reduced population needs more reduction, they will be sent aging to parallel genetic algorithm and to map reduce. (i.e. repeat steps from 2 to 7 to achieve the desired sets)

The end user can specify the code for the PGA to generate the random initial population. Meanwhile the end user can manage the overall execution of Genetic Algorithm.[11]

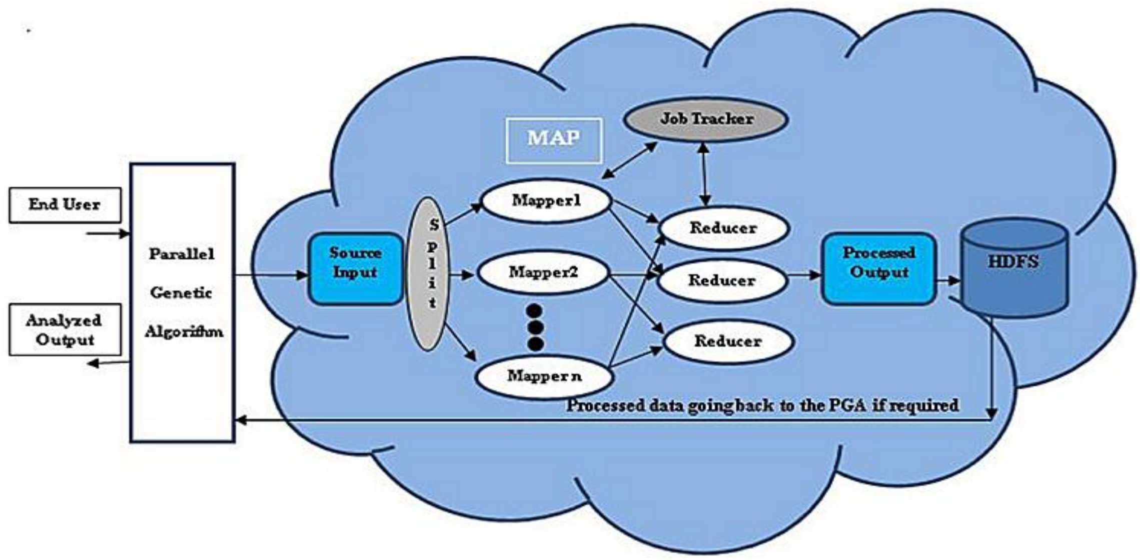

Fig 2: Suggestive Model for PGA with Hadoop for Medical Image Processing

5.2.1 Source Input

The source input gets the initial population and gives them to the split mechanism for further assigning of them to the mappers available.

5.2.2 Split

Split mechanism is to split input image data population to prepare the input for the mappers. Depends on the number of mappers available the number of input divided set is dynamically processed.

5.2.3 Mapper

Once the split does its splitting process Mapper receives input and carries out its task on the received input set in a parallel and independent way. Mapper is accountable to evaluate the corresponding fitness value of the received input data i.e. chromosomes. After review is completed, each *Mapper* produces a new pair <key, value>, where value is a pair <chromosome, fitness value>.

5.2.4 Job Tracker

The Job Tracker module in this phase is responsible to collect the outputs produced by the Mappers that will constitute the input for Reducer. The Job Tracker module is accountable not only to coordinate with the assignment of Reducers based on the yield from Mappers but also taking care also of the load balancing aspects with respect to the Reducers.

Once the new children are generated from the Reducers, Job Tracker module alerts Parallel Genetic Algorithm to restart the computation by invoking the MapReduce Job. The Master module assigns the Reducers to the mappers according to the new key it has received.

5.2.5 Reducer

Once the Mapper manipulates the data and produces the key and value, they are sent to the reducer. Key is the one generated newly and the corresponding chromosome and its fitness value are also together for the reducers.

Since Medical Image Processing may require different attributes, there may be many Reducers based on the given criteria. Since the PGAs are working based on survival of the fittest, now the Reducers can select the fittest ones. From these fittest ones, by applying the crossover and mutation operators, Reducers can generate new children. They can be evaluated through the next MapReduce flow.

5.2.6 HDFS

Hadoop Distributed File System is a scalable, fault-tolerant, distributed storage system that works closely with MapReduce. Because of its high availability one can be sure that he/she can achieve the goal without any inhibitions.

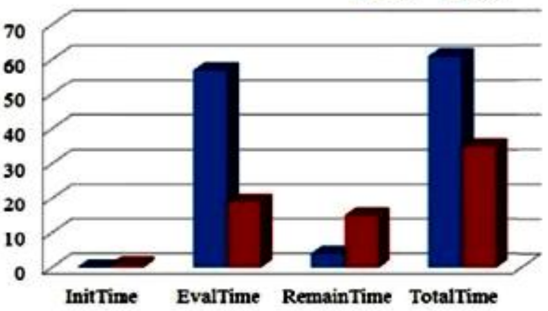

Fig 3: Execution time performance of SGA and PGA

Sample evaluation of Sequential Genetic Algorithms (SGA) and PGA in the Hadoop environment is given in Fig 3.

From the above figure, one can understand that there is tremendous change in the execution time while using the PGA with Hadoop. 57% of TotalTime is highly reduced with PGA compared to SGA. Each map initialization time (i.e., InitTime) is much less than the RemainTime, due to data management in reduce phase. Even this test is not dealt with medical images, it can be concluded that there is surely a drastic decrease in the execution time when incorporating PGA and Hadoop for Medical Image Processing.

6. FUTURE RESEARCH DIRECTIONS

The analysis shows the effectiveness of parallel genetic algorithm in the Hadoop setup. Recent times it is a great challenge for a doctor to diagnose more than getting an image. If a data can speak the true image of the human brain, it will be a great help for a doctor. Image processing is an important key to the future since it will enable the linking images to other types of data for mining. So research on Medical Image Processing takes a lead and it can contribute a lot to this.

Since people are moving towards mobile applications, composing the image processing easier, faster and accurate may help more for the developers. Researchers can improve the efficiency of the Hadoop setup and help the PGA to run even faster to yield the best result. Developing applications for various platforms in the mobile and cloud environment also can be a one of the best research leads from this point.

7. CONCLUSION

PGAs have already been used successfully in MIP through parallelization and Hadoop also effectively exhibits parallel processing with map reduce. When one look into these features of PGAs and Hadoop it is sure that better efficiency can be achieved in the execution of PGAs.

There are challenges remain to its implementation in medical image processing. May be the foremost challenge is the amount of data scattered in various places and in different formats. The authenticity of the data may be the other side of the challenge. Overcoming these barriers with the help of researchers and the respective people will definitely lead to a better solution for the future.

8. ACKNOWLEDGMENTS

It is my privilege to thank the expert who has contributed towards knowledge input to write this paper. I would like to thank my guide Dr. Pethuru Raj, IBM Global Cloud Center of Excellence, Bangalore for providing various online resources. The immense experience of Dr. Pethuru Raj in developing and implementing Cloud applications especially with the medical image processing applications is an exposure for me to study and take a lead for this paper.

9. REFERENCES

[1] Di Geronimo, L. Ferrucci, F. Murolo, A. and Sarro, F. 2012. A Parallel Genetic Algorithm Based on Hadoop MapReduce for the Automatic Generation of JUnit Test Suites. IEEE Fifth International Conference on Software Testing, Verification and Validation (ICST). (April 2012), 785-793.

[2] Muhammad, A. Bargiela, A. King, G. 1997. “Fine-Grained Parallel Genetic Algorithm: A Stochastic Optimisation Method,” The First World Congress on Systems Simulation. (1997), 199-203

[3] Apache Hadoop 2.6.0. http://hadoop.apache.org/docs/current. (November 2014)

[4] The basic genetic algorithm. http://www.edc.ncl.ac.uk/highlight/rhjanuary2007g01.php. November 2014

[5] Bulu, H. and Alpkocak, A. 2007. “Comparison of 3D Segmentation Algorithms for Medical Imaging,” Twentieth IEEE International Symposium on Computer-Based Medical Systems. (June 2007), 269-274.

[6] White, T. Hadoop: the Definitive Guide (2nd Edition) [M]. O’Reilly Media, 2010.

[7] Yang Song. Alatorre, G. Mandagere, N. Singh, A. 2013. Storage Mining: Where IT Management Meets Big Data Analytics, Big Data (BigData Congress), IEEE International Congress.

[8] Dipali, M. Joshi, Rana, N. K. and Misra, V. M. “Classification of Brain Cancer Using Artificial Neural Network” 2010 2nd International Conference on Electronic Computer Technology (ICECT 2010).

[9] Divya, K. Utkarsha, S. and Paridhi, S. 2013. “Medical Image Segmentation using Genetic Algorithm,” International Journal of Computer Applications, (Nov 2013), 9-15.

[10] Elavarasi, K. and Jayanthy, A. K. “Soft sensor based brain tumor detection using CT-MRI,” International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 10, 2013.

[11] Angel, K. S. and Jayakumari, J. 2011. Automatic Detection of Brain Tumor based on Magnetic Resonance Image using CAD System with watershed segmentation. In Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies(ICSCCN 2011).

[12] Mantas, P. And Andrius, U. 2007. “A Survey of Genetic Algorithms Applications For Image Enhancement And Segmentation,” Information Technology And Control, Vol. 36, 278-284.

[13] Peter, A. 2014. “Leveraging Big Data Analytics and Hadoop in Developing India’s Healthcare Services,” International Journal of Computer Applications, (March 2014), 44-50.

[14] K.Selvanayaki, Dr. M. Karnan.” CAD System for Automatic Detection of Brain Tumor through Magnetic Resonance Image-A Review,” International journal of Engineering Science and Technology, 2010, 5890-5901.

[15] Francis, G. Parallel Genetic Algorithm, Chapter 3. http://whitedwarf.org/metcalfe/node8.htm. 2014

[16] Intelligent Control Techniques in Mechatronics - Genetic algorithm. 3.3.10 http://www.ro.feri.uni-mb.si/predmeti/int_reg/Predavanja/Eng/3.Genetic%20algorithm/_16.html. November 2014