Genetic Algorithms for Staging Cervical Cancer

Sookham R.P Singh
M.Tech-Computer Science & Engineering
Guru Nanak Dev University
Amritsar

Rajinder Singh Virk
Associate Professor, Dept of CSE
Guru Nanak Dev University
Amritsar

ABSTRACT
The paper describes an application of engineering technology in the field of medical science i.e. a system for classifying the different stages of cervical cancer using an evolutionary computing technique, genetic algorithms. The proper staging of cervical cancer is the most important factor in selecting the right treatment plan. As there is a large amount of information to be processed during the course of the diagnosis of cancer, so the use of an efficient search and optimization technique would ease the process and decrease the efforts.

Keywords
Evolutionary, genetic algorithms, feature extraction

1. INTRODUCTION
The principle of evolution is the primary concept of biology, linking every organism together in a historical chain of events. Those individuals that are better are more likely to survive and propagate their genetic material [1].

The biologically motivated computing activities have evolved over years. The first has grown into the field of neural networks, the second into machine learning, and the third into what is now called "evolutionary computation," of which genetic algorithms are the most prominent example. Evolutionary systems work on the idea that evolution could be used as an optimization tool for engineering problems. The idea in all these systems was to evolve a population of candidate solutions to a given problem, using operators inspired by natural genetic variation and natural selection [2].

Genetic algorithms are randomized search optimization techniques guided by the principles of evolution of natural genetics. They are efficient, adaptive and robust search processes producing near-optimal solutions and having a large amount of implicit parallelism. Genetic algorithm is a global search technique useful for complex optimization problems where the number of parameters is large and difficult to obtain [3].

Cancer refers to a class of diseases in which a cell or a group of cells divide and replicate uncontrollably, intrude into adjacent cells and tissues and ultimately spread to other parts of the body than the location at which they arose [4]. The world wide occurrence of cervical cancer cases show that only 20% of these cases are found in developed nations while 80% of the cases are found in the developing countries that include India [5]. Cervical cancer is a malignant neoplasm arising from cells originating in the cervix uteri.

2. RELATED WORK
Cervical cancer is one of the most common cancers among women worldwide. In the research paper by Ambika Satija [6], the author briefs about the statistics of cervical cancer in India. In the paper by G. Jayalalitha and R. Uthayukumar, a method of grading cervical cancer images according to the cell formation of tissues has been presented [7].In the paper by Catherine Todd and Rahmdwati, G.Naghdy [8], an algorithm for computer assisted classification of cervical cancer using digitized histology images of biopsies has been presented.

Another paper that focuses on the use of histology images for the classification of cervical cancer was proposed by Montse Ross and Rahmadwati [9].

The research paper by S.Allwin and S.Pradeep Kumar [10], proposes an approach to classify the various malignancies in cervical cyto images using textural properties of the cervical cyto image. In the paper by C. Balleyguier et al. [11], guidelines for the staging and follow-up of patients with uterine cervical cancer have been provided. The paper by Pratibha Bajpai et al. [12], gives a detailed study of genetic algorithms. Yet another paper by Tom V. Mathew [13], describes the canonical genetic algorithm, its basic working principle, coding, operators, etc.

The paper by Pengfei Guo et al. [14] presents a hybrid genetic algorithm for optimization problems. The paper by Keshavamurthy B.N et al. [15], proposes an approach that improves the evolutionary technique such as genetic algorithm by improving the fitness function parameters. The research work by Mohd Saberi Mohamad et al. [16], introduces a new algorithm of a hybrid of genetic algorithm and support vector machine for gene selection and classification task.

The paper by Mausumi Bharadwaj et al., shows a novel association of cell cycle regulatory gene, cyclin D1 (CCND1) variation with cervical cancer cases in Indians [17]. In the paper by Andrew K. et al. [18], an integrated genetic algorithm for gene expression data analysis has been proposed.

The research work by Fadzil Ahmad et al. [19], uses GA to simultaneously select significant features as input to ANN. The paper by Pabitra Mitra et al. [20], describes a way of designing a hybrid decision support system in soft computing paradigm for detecting the different stages of cervical cancer.

3. EVOLUTIONARY COMPUTING
Evolutionary computing is the study of robust search algorithms based on the principles of evolution.

The principle of evolution is the primary unifying concept of biology, linking every organism together in a historical chain of events. Over many generations random variation and natural selection shape the behaviors of individuals and species to fit the demands of their surroundings.

Three types of evolutionary computing techniques are widely reported recently. These are Genetic Algorithms (GAs),

Genetic Programming (GP) and Evolutionary Algorithms (EAs) [21].

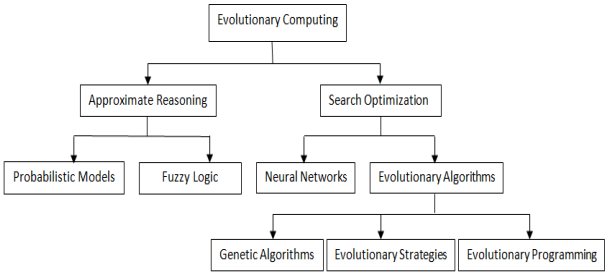

Figure 1. Evolutionary computing techniques

3.1 Genetic algorithms

The genetic algorithms (GA's) are randomized search optimization techniques guided by the principles of evolution of natural genetics. They are search heuristic that is routinely used to generate useful solutions to optimization and search problems. It generates solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover [12].

Genetic algorithms encode the decision variables of a search problem into finite-length strings of alphabets of certain cardinality. The strings which are candidate solutions to the search problem are referred to as chromosomes, the alphabets are referred to as genes and the values of genes are called alleles.

To evolve good solutions and to implement natural selection, we need a measure for distinguishing good solutions from bad solutions. The measure could be an objective function that is a mathematical model or a computer simulation, or it can be a subjective function where humans choose better solutions over worse ones. In essence, the fitness measure must determine a candidate solution's relative fitness, which will subsequently be used by the GA to guide the evolution of good solutions [22].

4. STAGING

Staging is the process of finding out how far the cancer has spread. Information from exams and diagnostic tests is used to determine the size of the tumor, how deeply the tumor has invaded tissues within and around the cervix, and the spread to lymph nodes or distant organs (metastasis). This is an important process because the stage of the cancer is the key factor in selecting the right treatment plan [23]. The stages of cervical cancer are:

4.1 Stage 0 (Carcinoma in situ)

In carcinoma in situ (stage 0), abnormal cells are found in the innermost lining of the cervix.

4.2 Stage I

In stage I, cancer is found in the cervix only.

a) Stage IA: A very small amount of cancer that can only be seen with a microscope is found in the tissues of the cervix. In this stage the cancer is not more than 3 mm deep and not more than 7 mm wide.

b) Stage IB: is divided into stages IB1 and IB2. In this stage the cancer is more than 5 mm deep and more than 7 mm wide or is larger than 4 cm.

4.3 Stage II

In stage II, cancer has spread beyond the cervix but not to the pelvic wall or to the lower third of the vagina.

a) Stage IIA: Cancer has spread beyond the cervix to the upper two thirds of the vagina but not to tissues around the uterus. In this stage the tumor can be seen without a microscope and is 4 centimeters or larger.

b) Stage IIB: Cancer has spread beyond the cervix to the tissues around the uterus.

4.4 Stage III

In stage III, cancer has spread to the lower third of the vagina, and/or to the pelvic wall, and/or has caused kidney problems.

a) Stage IIIA: Cancer has spread to the lower third of the vagina but not to the pelvic wall.

b) Stage IIIB: Cancer has spread to the pelvic wall; and/or the tumor has become large enough to block the ureters. This blockage can cause the kidneys to enlarge or stop working.

4.5 Stage IV

In stage IV, cancer has spread to the bladder, rectum, or other parts of the body.

a) Stage IVA: In this stage the cancer has spread to nearby organs, such as the bladder or rectum.

b) Stage IV B: In this stage the cancer has spread to other parts of the body, such as the liver, lungs, bones, or distant lymph nodes [24].

5. METHODOLOGY

Genetic Algorithms are highly parallel and adaptive search processes based on the principles of natural selection. In this work GA's are used for searching and optimizing the solutions. The methodology followed uses genetic algorithms for classifying the stages of cervical cancer by extracting some rules. The genetic operators like selection, mutation and crossover are used to generate and evolve a population of all possible solutions. The population in this problem consists of chromosome strings formed by a combination of the parameters or symptoms of the cervical cancer data.

5.1 Design of proposed GA

The GA involves following procedures- encoding of problem parameters in form of binary strings, application of genetic operators like crossover and mutation, selection of individuals based on some objective function to create a new population.

5.1.1 Chromosomal Representation

A chromosome is a set of genes, which contains the solution. A method is needed to encode potential solutions to that problem in a form so that computer can process it. For this problem hybrid encoding is chosen as the parameters contained in the cervical cancer patient's data are in a mixed format. Each of the parameter is encoded into a string of 11 bit length. The structure of the chromosome is as follows:

H	N	C	M	U	S	P	P	A	S	D

Table 1. Chromosomal representation

Bit Number	Name	Type	Alleles
Bit 1: H	Histology	character	E, S, C
Bit 2: N	Node PET	integer	0,1
Bit 3: C	Clinical diameter	double	1.0-8.0
Bit 4: M	MRI volume	double	0.0-300.0
Bit 5: U	Uterine body	integer	0,1
Bit 6: S	Status	integer	0,1
Bit 7: P	Relation Primary	integer	0,1
Bit 8: P	Relation Pelvic	integer	0,1
Bit 9: A	Relation Abdomen	integer	0,1
Bit 10: S	Relation Supraclav	integer	0,1
Bit 11: D	Relation Distant	Integer	0,1

The strings generated in this pattern form the initial population for the problem. For example:

Chromosome -> S13.92.20100011
Chromosome -> C09.02.90111100

5.1.2 Fitness function

Fitness is the value assigned to an individual based on how far or close an individual is from the solution; greater the fitness value better the solution it contains.

Fitness function is a function that assigns fitness value to an individual. It is problem specific. Fitness function qualifies the optimality of a solution (chromosome) so that a particular solution may be ranked against all other solutions. The function depicts the closeness of a given solution to the desired results.

In this problem an objective function is derived based on the data of cervical cancer patients and its parameters. The fitness function can be represented as below:

$$\text{Fitness} = \text{Normalization Constant} - \left(\text{Clindiam}_{j\text{-}i} + \text{MRIvol}_{j\text{-}i} + \sum_{n=1}^{n=11*} 1 \right)$$
$$\&\& \; n!=3,4$$

where,

i: input string given by user based on parameters
j: random string generated by genetic algorithm
* i[n] != j[n]
Clindiam: the clinical diameter derived from the patient's data
MRIvol: the MRI volume derived from the patient's data

This fitness function depicts the difference between the string provided as input by the user in form of the parameters mentioned above and the chromosome string generated by the genetic algorithm. The string with the highest difference is considered to be fittest in this case and selected.

5.1.3 Selection

Selection is the first operator applied on the population generated. From the population, the chromosomes are selected to be parents to crossover and produce offspring.

In this problem, once the population of chromosomes has been generated, the selection of parents has been done based on a fitness function mentioned above.
The Roulette wheel selection method or the fitness proportionate selection method has been used for this problem. The fitness function has been derived on the basis of the parameters of data of cervical cancer patients. This function selects two fittest parents out of the population of chromosomes generated initially. The selected parents are then passed on to the crossover operator.

5.1.4 Crossover Operator

Crossover is a genetic operator that combines two chromosomes (parents) to produce a new chromosome (offspring). Crossover occurs during evolution according to a user-defined probability. Single point crossover is adopted for this problem since the string length is not very large. In case of large string length, single point crossover would have little effectiveness. Multiple-point crossover would be adopted to ensure a high probability for only one crossover point occurring within a word.
The crossover probability for this problem is fixed at 0.8.

5.1.5 Mutation Operator

After crossover is performed, mutation takes place. Mutation alters one or more gene values in a chromosome from its initial state. This can result in entirely new gene values being added to the gene pool. With the new gene values, the genetic algorithm may be able to arrive at better solution than was previously possible. Mutation is an important part of the genetic search, helps to prevent the population from stagnating at any local optima. Mutation is intended to prevent the search falling into a local optimum of the state space.
Mutation occurs during evolution according to a user-definable mutation probability.
The mutation probability for this problem is set to be 0.6.

5.2 Classification rules

The data in this case consists of a set of 239 cervical cancer patient cases that have been obtained from the database of International Gynaecologic Cancer Society (IGCS) [25]. There are four classes corresponding to the Stages I-IV of the cancer, each containing some patient cases.

The classification rules have been extracted on the basis of the occurrence of parameters and their possible values. These rules have been used to detect the stages of cervical cancer.

Table 2.Classification rules extracted

Stage	Rules
1a	SCC Λ Clin(0.0-1.0) Λ MRI(0.0-30.0)
1b	SCC Λ Clin(1.0-2.0) Λ MRI(30.0-50.0)
2a	SCC,Endo Λ Clin(2.0-3.0) Λ MRI(50.0-70.0)
2b	SCC,Endo Λ Clin(3.0-4.0) Λ MRI(70.0-90.0)
3a	SCC,Endo Λ Clin(4.0-5.0) Λ MRI(90.0-100.0)
3b	SCC,Endo Λ Clin(5.0-6.0) Λ MRI(100.0-150.0)
4a	SCC,Endo Λ Clin(6.0-7.0) Λ MRI(150.0-200.0)
4b	SCC,Endo,Clear Λ Clin(7.0-8.0) Λ MRI(200.0-300.0)

5.3 Outline of GA

The flow chart below, Fig. 1, shows the rough working of the proposed system. The system classifies the stages of the

cervical cancer and recognizes similar diagnostic cases in the data. This has been achieved by exploiting the randomized search property of genetic algorithms. The data of the cervical cancer patients is provided as input and based on some classification rules extracted the stage of the cancer is detected. This input string provided by the user along with a randomly generated chromosome is selected from the population for crossover and mutation. The best chromosome is picked up and the similar cases are shown from the data of patients for diagnostic recommendation.

The performance of the system with time i.e. the changes in crossover and mutation probability with time and the effect of probability of crossover and mutation on the number of generations, as well as the comparison of genetic algorithms with deterministic search algorithms has been studied.

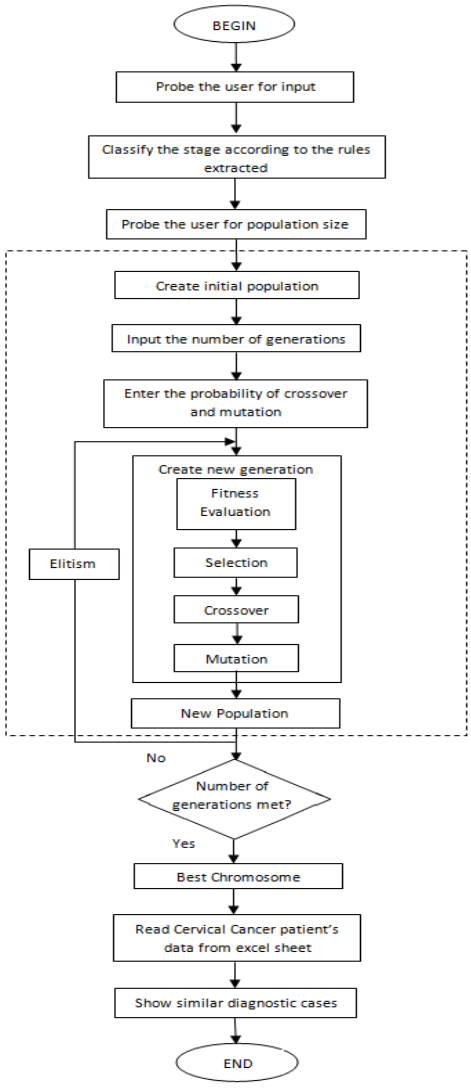

Figure 2. Flow chart

The crossover probability is analyzed, by keeping the population size, number of generations and the mutation probability fixed at a certain value. The Figure 3 shows that the probability of crossover, pc, gives better results at a value pc= 0.6 to pc= 0.8. Hence for this problem, a crossover probability of 0.8 has been chosen for better results.

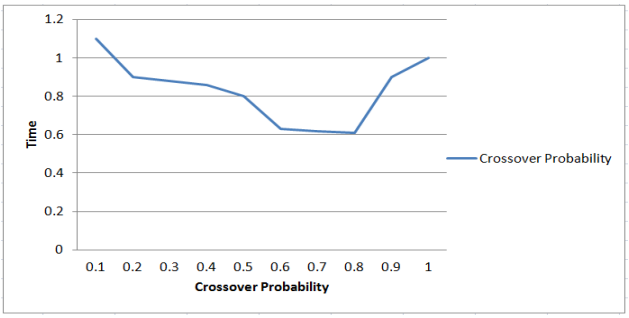

Figure 3. Crossover Probability Graph

The probability of mutation has been analysed by keeping the size of the population, the number of generations and the crossover probability fixed at a certain value. The analysis graphical shows that, the probability of mutation, pm, gives optimal results at a value pm= 0.4 to pm= 0.6. Thus for this problem, the mutation probability is chosen to be 0.6

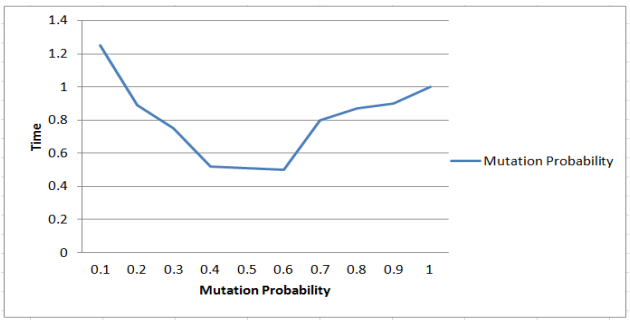

Figure 4. Mutation Probability Graph

The effect on the number of generations has been analyzed with respect to the probability of crossover and mutation, by keeping the population size fixed and varying the probability of crossover and mutation from 0.1 to 1. The Figure 5 shows that the near optimal results are achieved at crossover probability, pc= 0.8 and mutation probability, pm= 0.6.

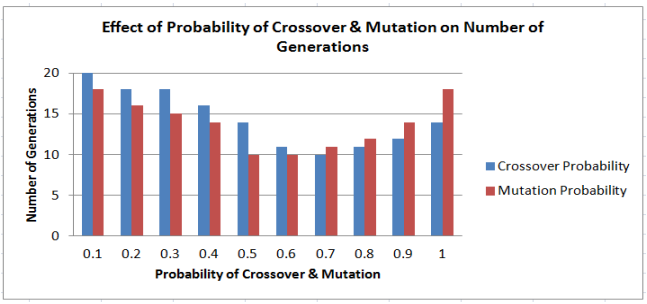

Figure 5. Effect on number of generations

A comparative analysis of the search using genetic algorithms and the other deterministic search algorithms has been shown in Figure 6. The graphical analysis shows that, the performance of genetic algorithms up to the data entries ≈ 250 increases slowly and steadily. The deterministic search algorithms give better results up to data entries < 250, as they take less time to search whereas, the genetic algorithms take comparatively more time to search.

However, at data entries > 250, the performance of genetic algorithms improves exponentially, as they show better results, taking less time to search. Whereas, the performance of deterministic search algorithms deteriorates after this point as they take much time to search for large data. Hence, for this

problem, where the data entries are near about 250, the genetic algorithms have been chosen to provide near optimal results.

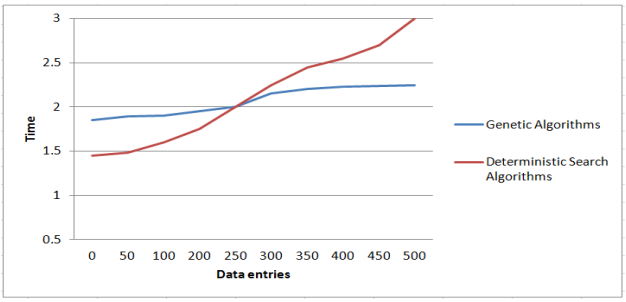

Figure 6. Comparison of GA with Deterministic Search

6. CONCLUSION

In the field of oncology, diagnosis and treatment require a large amount of information to be processed. A computerized program developed for classifying the stages and recognizing the similar diagnostic cases, would make the task of a practicing oncologist much easier. The proposed system makes use of the genetic algorithms for the staging of the cancer of uterine. The results show that genetic algorithms work efficiently in processing the large amount of information of the cervical cancer patient's provided as input and then searching similar diagnostic cases from the large amount of data.

7. REFERENCES

[1] David B. Fogel, "What is evolutionary computation?", Natural Selection Inc., IEEE Spectrum, Feburary 2000, page 26

[2] Mitchell Melanie, Introduction to Genetic Algorithms, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England, Fifth Printing, 1999.

[3] David E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Pearson Education, Dorling Kindersley(India) Pvt. Ltd., Eighth impression, 2012

[4] "Cevical Cancer Overview", American Cancer Society,http://www.cancer.org/acs/groups/cid/documents/webcontent/003042-pdf.pdf

[5] K.Jayant, R.S Rao, "Improved stage at diagnosis of Cervical cancer with increased cancer awareness in rural Indian population", International Journal Cancer,1995, Vol.63, pp 161-163

[6] Ambika Satija, "Cervical cancer in India", South Asia Centre for Chronic Disease, 2009.

[7] G. Jayalalitha and R. Uthayukumar, "Recognition of Cervical cancer based on Fractal Dimension", International Conference on Advances in Recent Technologies in Communication and Computing, 2009

[8] Catherine Todd and Rahmdwati, G.Naghdy, "Cervical Cancer Classification Using Gabor Filters", First IEEE International Conference on Healthcare and Informatics, Imaging and Systems Biology, 2011

[9] Montse Ross and Rahmadwati, "Classification Cervical Cancer Using Histology Images", Second international Conference on Computer Engineering and Application, 2010

[10] S.Allwin and S.Pradeep Kumar, "Classification of stages of Maligancies using Textron signature of Cervical Cyto Image", Computational Intelligence and Computing Research (ICCIC), 2010

[11] C. Balleyguier et al., "Staging of uterine cervical cancer with MRI", Journal on European Radiology, 2011

[12] Pratibha Bajpai et al., "Genetic Algorithm- an Approach to Solve Global Optimization Problems", Indian Journal of Computer Science and Engineering, 2010

[13] Tom V. Mathew, "Genetic Algorithm", Journal of Computing and Engineering, 2005

[14] Pengfei Guo et al., "Enhanced Genetic Algorithms for Optimization Design", 3rd International Conference on Biomedical Engineering and Informatics, 2010

[15] Keshavamurthy B.N et al., "Improved Genetic Algorithm Based Classification", International Journal of Computer Science and Informatics, 2010

[16] Mohd Saberi Mohamad et al., "Gene selection from Gene Expression Data using Genetic Algorithm for Cancer Classification", Third International Conference on Artificial Intelligence in Engineering and Technology, 2009

[17] Mausumi Bharadwaj et al., "Genetic Basis of HPV meditated Cervical Cancer in Indian Women", International Conference on Systems in Medicine and Biology, 2010

[18] Andrew K. et al., "Cancer gene-search with data mining and Genetic algorithms", Journal of Computer Science and Medicine, 2009

[19] Fadzil Ahmad et al., "Genetic Algorithm –Artificial Neural Network (GA-ANN) Hybrid Intelligence for Cancer Diagnosis", Second International Conference on Computational intelligence, Communication Systems and Network, 2010

[20] Pabitra Mitra et al., "Staging of Cervical Cancer with Soft Computing", IEEE Transactions on Biomedical Engineering, 2000

[21] "Introduction to evolutionary computing techniques", Electronic Technology Directions to the Year 2000, 1995. Proceedings, Publication Year: 1995 , Page(s): 122 – 127, IEEE Conference Publications

[22] Kumara Sastry, David Goldberg et al., "Genetic Algorithms", Springer US, 2005, Pages: 97-125

[23] Manik Sharma. "Role and Working of Genetic Algorithm in Computer Science". International Journal of Computer Applications and Information Technology (IJCAIT). Volume 2, Issue 1.

[24] http://www.cancer.org/cancer/cervicalcancer/detailedguide/cervical-cancer-staged

[25] http://www.cancer.gov/cancertopics/pdq/treatment/cervical/Patient/page2

[26] www.igcs.org/professional/Education/treatmentResources/CervicalCaDB.html

[27] Manik Sharma, Gurvinder Singh, Gurdev Singh, Rajinder Singh. "Analysis of DSS Queries in Distributed Database System Using Exhaustive and Genetic Approach". International Journal of Advanced Computing, Vol.36, Issue.2, April 2013.

[28] Manik Sharma, Gurdev Singh and Harsimran Kaur. "A Study of BNP Parallel Task Scheduling Algorithms Metric's For Distributed Database System". International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.1, January 2012

[29] Sharma, Manik, and Gurdev Singh. "Analysis of Joins and Semi-joins in Centralized and Distributed Database Queries."*2012 International Conference on Computing Sciences*. IEEE, 2012.

[30] Sharma, Manik, Gurdev Singh, and Rajinder Virk. "Analysis of Joins and Semi Joins in a Distributed Database Queries."*International Journal of Computer Applications* 49.16 (2012).

[31] Sharma, Manik, Gurdev Singh, and Rajinder Virk. "Analysis of DSS queries in distributed database system." *International Journal of Networking and Parallel Computing* 1.3 (2012).

[32] Sharma, Manik. "Role and Working of Genetic Algorithm in Computer Science." *International Journal of Computer Applications and Information Technology (IJCAIT)* 2.1 (2013).