Clinical Cancer Investigation Journal-ISSN: 2278-0513

Detection of Types of Cancer of Unknown Primary in Big Data Using Map-Reduce Model

Abstract

This study aims to detect types of Cancer of unknown primary in big data using the MapReduce model. In this research, we want to integrate two calculation methods, i.e., using the K-means clustering algorithm in the parallel programming (MapReduce) text, so that we can acceptably achieve the minimum difference between the actual number of classes in the dataset and the number of recycled clusters. The Rand index is the validation criterion. This research used standard Euclidean Similarity (EZ0) and Cosine Similarity (C) criteria. A total of 18 Rand indices were obtained by running the algorithm on 18 datasets, analyzing the number of samples in each cluster, and comparing them with the real classes in each dataset. Accordingly, 18 Rand indices were obtained, and the result of this operation was monitored with the Euclidean similarity. This result was extracted by averaging the 18 mentioned indices between the FMG and K-Means method compared to the output.

Keywords: *Clustering, K_Means, MapReduce, Gene Expression Data*

Vahid Reza Jahanpour[1*], Mohammad Amin Irandoost[2]

1. Master's in computer, computer department, Faculty of Engineering, Islamic Azad University, Hamedan branch, Hamedan, Iran
vr.jahanpour@iauh.ac.ir
vr.jahanpour@gmail.com
2. Assistant professor of Computer Engineering, Islamic Azad University Hamedan Branch, Hamedan, Iran
Email: aminirandoost@gmail.com
aminirandoost@iauh.ac.ir
corresponding author: Vahid Reza Jahanpour

Introduction

Now cancer is one of the important and main challenges of health and treatment in Iran and worldwide. It is the second cause of death after Cardiovascular diseases (CVDs). In developing countries, cancer is one of the most important health problems, and its trend is growing [1]. Timely diagnosis and proper treatment of this disease improve patients' recovery rate and survival range. Studies show that the use of modern computer technologies, such as image processing mechanisms, has been successful in diagnosing and classifying cancers, while cancer diagnosis has been based on interventions such as surgery, radiotherapy, and chemotherapy [2]. Cancer of Unknown Primary (CUP) is a classification for tumors when cancer metastases are seen, but the origin of the primary tumor is unknown. Some 3 to 5% of all newly diagnosed cancers are CUPs, and in many cases (80-85%), there is a poor prognosis. In classical medical methods, CUP is somewhat challenging because the primary tumor remains a mystery in up to 70% of CUP cases [3].

Studying gene expression profiles of cells and tissues is an important tool for detection in medicine. Microarray experiments characterize genomic expression changes in health and disease [4]. Large companies have accumulated a huge amount of data in parallel with the rapid development of network technology and information technology. However, a large amount of valuable knowledge is hidden, which is not directly proportional to data growth, especially when this data increases. Therefore, researchers focus on how to find valuable knowledge as well as how to better use the information in the future, and hence Data mining has emerged in the last years. The use of clustering techniques has been emphasized by many scientific societies to discover cancer subspecies. Microarray technology makes it possible to measure the molecular Signatures of cancer cells. The K-means algorithm was developed by McQueen and James [5] in 1967. The K-means ***technique*** is a popular method for analyzing gene expression data [6]. It is an iterative algorithm for classification that optimizes the best fit between clusters and their representation using predefined clusters [5]. Jordan and Weiss [6] introduced the Spectral clustering technique. Finally, this clustering also uses an algorithm such as K-Means, but it creates a series of changes in the data structure and a change in the way you look at data. In [7], the authors have described hierarchical methods. In these algorithms, the number of clusters is not assumed to be a specific value and can be from one to n clusters. We obtain the number of different clustering states by running an algorithm and showing the clusters according to the user's needs.

Although the K-means method is known as one of the most popular and widely used clustering techniques, experience shows that clustering big data using classical methods alone is not very efficient. In [8], the parallelization solution of the K-means method is discussed in the framework of the MapReduce model. The main idea is that in the K-means method, distance calculations between one object with centers are not related to distance calculations between other objects with their desired centers. This can lead us to parallel computing because the parallelization of processes requires them to be independent of each other [9].

Many clustering algorithms have been proposed to analyze gene expression data. That is, we provide a method using the K-means clustering algorithm in the parallel programming (MapReduce) text so that we can acceptably minimize the difference between the actual number of classes in the dataset and find the number of the recovered cluster. The Rand index is a validation measure.

Material and Method

Recommended approach

```
Procedure K-MeansMap([μ1 , μ2 ,…, μn], xi)
Begin
Index=0;
min=dist(μ1, xi);
For Count-μ To k DO
Begin
 If dist((μCount-μ , xi) < min Then
 Begin
        Min=Dist(μCount-μ, xi)//
      This function calculates the Euclidean or cosine distance of each data point
      from the cluster center
        Index=k;
End;
End;
Emit(μIndex , xi);
End
```

Manhattan, Euclidean, and Minkowski metrics are used,

```
Procedure K_MeansReduce(j, Cluster_j: [x1, x2, …])
Begin
      Sum[]=0  // an array that stores the sum of all characteristics of each sample
      count=0  // The total number of samples in the cluster

    For i=1 To TotalRow(Cluster_j) Do
     Sum[]+=x[i];          // here, we add the sum of the values of the corresponding
                           characteristics of the samples for averaging
    Emit(j, Sum[]/ TotalRow(Cluster_j)
End;
```

considering the numerical nature of the special characteristics of the research samples. First, it should be noted that the problem of clustering cancer gene expression data (tissues) is

```
Function Dist(x1(p1, p2...pn),  x2(p1, p2...pn) )
Begin
      For i=1 To n Do
        Dist=Sqrt((x1p1 - x2p1)^2+(x1p2 - x2p2)^2+…+(x1pn - x2pn)^2
    Return(Dist);
End;
```

very different from gene clustering before describing the proposed approach. In texture clustering, tens or hundreds of elements (textures) must be classified, and each of these data elements is described by thousands of genes. In contrast, in gene clustering, there are a lot of data elements (genes), which are described by a small number of different modes. Accordingly, the clustering of a few elements (tissues) with large dimensions is not the same as the clustering of a lot of elements (genes) with small dimensions [10]. In the present study, the latter (gene clustering) is used. In the first step, which is called the Classify Step, we assign each Data Point to a cluster. Then we have a step called the Recenter Phase, where we use all the data points assigned to a given cluster center to ***re-***update the center of that cluster.

1. Implementation of K-Means problem steps in MapReduce

In the first step, parallel data operations are important. That is, for each data point, whenever a cluster center is determined, then it is possible to independently assign that data point to a cluster center. It never depends on other data points, especially since the Mapper function is supposed to produce the pair (data, cluster center). In this pair, cluster center and data value are used as Key and Value, respectively. Then in the Recentering phase, we see that this is the step where we perform the aggregation operation. This aggregation is also performed independently between different cluster representatives (different keys). Another important point is that each of the machines must have the characteristics of all the centers of the clusters in the Map phase to calculate the minimum distance between each of the data points and the centers of the clusters. This was observed in the simulations.

2. Map phase

In the K-Means algorithm, the Map phase corresponds to the classification phase; That is, the Map phase is executed on every x member of the dataset, and finally, every x data point is assigned to the Nearest Cluster center (μ). In the Map phase, whose code fragment can be seen in Figure (1-a), all cluster centers as input parameters are sent to it. We find the x distance with each of the centers μi and find the μi with the smallest distance [11-15]. The output of the map phase is the pairs (Value, Key) produced by the emit function, which represents the assignment of each dataset member to a cluster representative (Figure 1-a).

a)

b)

c)

Figure 1: Map phase algorithm in K_Means (a), Reduce phase algorithm in K_Means (b), and Implementation of Euclidean distance algorithm between two points (c)

3. Reduce phase

The Reduce phase corresponds to the second phase of K-Means, i.e., the Recentering phase. The reducer takes a given class label (SameKey) along with the values list specified by that Key (Figure 1-b) and returns a pair (value, key) as output where Key is equal to the cluster label (here j), and value is equal to the center (Mean) of the new cluster.

As the Map algorithm shows, calculating the distance between samples is the task of the Dist function (Figure 1-b). The code of this function can be seen in Figure (1-c), and the calculation distance is based on the Euclidean **relations** and also on all of the sample characteristics. In this function, p1, p2...pn are the sample xi characteristics.

Evaluation Index

The evaluation of the structure of the retrieved clusters was done by the Rand index and by comparing the real classes of texture samples with the assigned clusters. Tests were done for each dataset. The proposed algorithm was implemented with each of the datasets, the Rand index was obtained for each dataset, and finally, its average was calculated to compare with previous works.

1. Definition of the Rand index

The Rand index was proposed by William M. Rand in 1971 and has been used by researchers as one of the most important metrics for evaluating clusters [12]. An Agreement criterion is needed to compare the clustering results with real **situations** (external). Let S={O1,…, On} be a set of n data types, and U={u1,..., uR} and V={v1,...vn} are two types of segmentation of data types, respectively, where U is the actual partitioning and V is our partitioning from the clustering. Since we consider that each gene belongs to only one class in real conditions and only one cluster (in our clustering); Accordingly, we have the following definitions:

a: the number of data pairs both in U and V

b: number of data pairs that are in U and not in V

c: number of data pairs that are in V and not in U

d: number of data pairs that are not in both U and V.

The Rand index is defined according to the following equation:

$$Rand\ Index = \frac{a+d}{a+b+c+d}$$

According to the above equation, the Rand index is between 0 and 1. The value of the above equation is equal to 1 when there are two partitions of the same size.

Results

In this study, the simulations were conducted in an environment with a CoreTM i7 CPU @ 2.3 GHz Intel® and using 6 GB of RAM. The operating system is Windows 7 Ultimate 64-bit. MATLAB has been used to simulate the Peer-to-Peer system, Database communication, and restructuring of data dissemination. We used our proposed algorithm on 18 datasets related to the introduction of genes to evaluate the accuracy. The characteristics of this dataset are listed in Table 1.

Table 1: Used Databases

No	Dataset name	Array type	Texture	Sample size	Class	Gene number
1	Armstrong-2002-v1	Affymetrix	Blood	72	2	1081
2	Armstrong-2002-v2	Affymetrix	Blood	72	3	2194
3	Bhattacharjee-2001	Affymetrix	Lung	203	5	1543
4	Chowdary-2006	Affymetrix	Breast, Colon	104	2	182
5	Dyrskjot-2003	Affymetrix	Bladder	40	3	1203
6	Golub-1999-v1	Affymetrix	Bone Marrow	72	2	1877
7	Golub-1999-v2	Affymetrix	Bone Marrow	72	3	1877
8	Gordon-2002	Affymetrix	Lung	181	2	1626
9	Laiho-2007	Affymetrix	Colon	37	2	2202
10	Nutt-2003-v1	Affymetrix	Brain	50	4	1377
11	Nutt-2003-v2	Affymetrix	Brain	28	2	1070
12	Nutt-2003-v3	Affymetrix	Brain	22	2	1152

13	Pomeroy-2002-v1	Affymetrix	Brain	34	2	857
14	Pomeroy-2002-v2	Affymetrix	Brain	42	5	1379
15	Ramaswamy-2001	Affymetrix	Multi-tissue	190	14	1363
16	Shipp-2002-v1	Affymetrix	Blood	77	2	798
17	Singh-2002	Affymetrix	Prostate	102	2	339
18	Su-2001	Affymetrix	Multi-tissue	174	10	1571

Disscussion

The basis of the comparison in the current research was the results [10] which can be seen in the diagram of Figure 2. As can be seen in the diagram on the left side of this figure, seven similarity criteria and seven clustering algorithms have been compared. The result of combining all the algorithms with the mentioned similarity criteria is shown by averaging the R index for clustering and true partition of classes [10] (Figure 2). The mentioned methods and similarity metrics are explained in Tables 1 and 2, respectively. Figure 2 shows the R index means for clustering and actual partition of classes [10]. For example, it can be seen that in terms of the cosine similarity criterion shown in the graph with the letter C, the K-Means algorithm with a 0.44 Rand index has a better performance than other algorithms with the same similarity criterion, while the Single-Link Hierarchical Clustering algorithm, abbreviated as SL, has the output minimum. Therefore, in other F algorithms, by combining impaired similarity criteria, it can be seen that the k-means output and Finite Mixed Gaussian (FMG) algorithms have better results than other methods due to their Rand index being closer to 1.

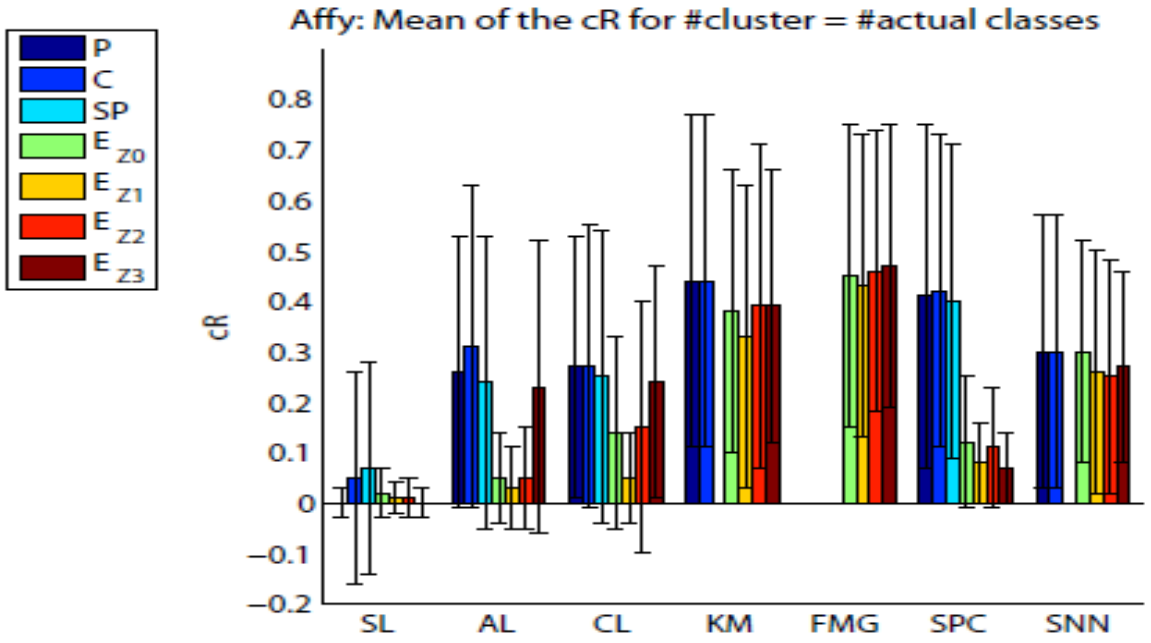


Figure 2: The R index-mean for clustering and actual partition of classes [7]

Table 2: Description of clustering algorithms and abbreviations (a), and description of similarity criteria (b)

a)

Type of clustering algorithm	Algorithm name	Acronyms
Hierarchical	Single Linkage	SL
	Complete Linkage	CL
	Average Linkage	AL
Partitioning	K-Means	KM
Partitioning	Finite Mixture of Gaussians	FMG
Partitioning	Spectral clustering	SPC
Partitioning	shared nearest neighbor-based	SNN

b)

Similarity criteria	Description	Acronyms
Pearson's Correlation coefficient	Pearson's correlation coefficient	SP
Cosine	Cosine similarity criterion	E_{Z0}
Spearman's correlation coefficient	Pearson's correlation coefficient between ranked data	E_{Z1}
Euclidean Distance Original	The original Euclidean distance	E_{Z2}
Euclidean Distance standardized	Standard Euclidean distance	E_{Z3}
Euclidean Distance scaled	Measured Euclidean distance	SP
Euclidean Distance ranked	Graded Euclidean distance	E_{Z0}

In this research, the standard Euclidean similarity criterion (EZ0) and cosine similarity criterion (C) were used. A total of 18 Rand indices were obtained by running the algorithm on 18 datasets (Table 2-a) and analyzing the number of samples belonging to each of the clusters and comparing them with the real classes in each dataset separately. The result of these operations with the Euclidean similarity criterion can be seen in Table 2-b. By averaging the above 18 results, we get a number that is between the FMG and K-Means method compared to the output of Figure 1-b. The results of the R index for each of the datasets by running the algorithm and using the Euclidean similarity criterion is shown in Table 3.

Table 3: R index for each of the datasets by running the algorithm and using the Euclidean similarity criterion

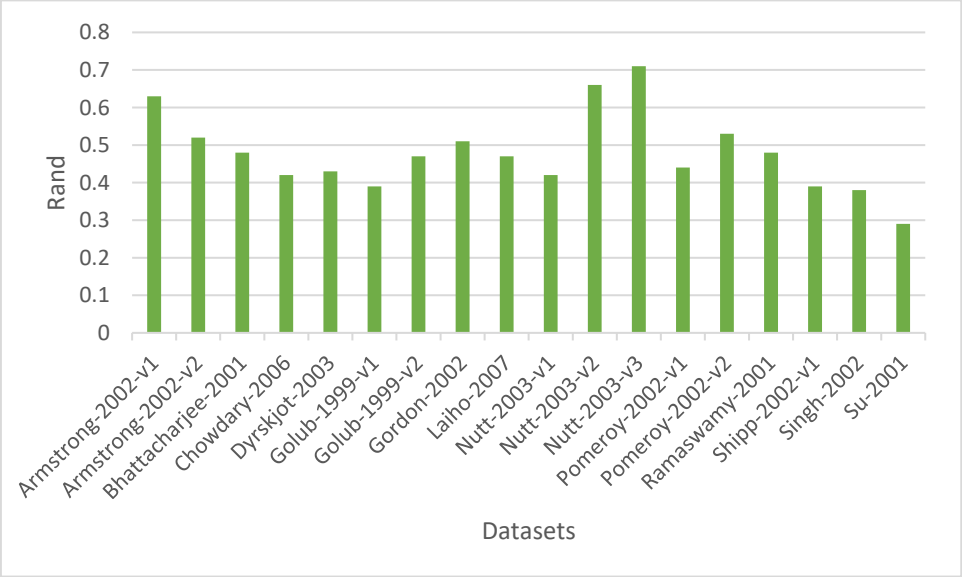
No	Dataset	Class number (actual cluster)	Gene number	Gene number in actual clustering	R index
1	Armstrong-2002-v1	2	1081	579،502	0.63
2	Armstrong-2002-v2	3	2194	658 ،805 ،731	0.52
3	Bhattacharjee-2001	5	1543	301 ،128 ،400 ،302 ،401	0.48
4	Chowdary-2006	2	182	77 ،105	0.42
5	Dyrskjot-2003	3	1203	191 ،611 ،401	0.43
6	Golub-1999-v1	2	1877	976 ،901	0.39
7	Golub-1999-v2	3	1877	554 ،815 ،508	0.47
8	Gordon-2002	2	1626	831 ،795	0.51
9	Laiho-2007	2	2202	1308 ،894	0.47
10	Nutt-2003-v1	4	1377	438 ،311 ،297 ،331	0.42
11	Nutt-2003-v2	2	1070	559،511	0.66
12	Nutt-2003-v3	2	1152	646 ،506	0.71
13	Pomeroy-2002-v1	2	857	459 ،398	0.44
14	Pomeroy-2002-v2	5	1379	459 ،314 ،189 ،219 ،205	0.53

15	Ramaswamy-2001	14	1363	،69 ،79 ،99 ،89 ،87 ،111 ،92 ،78 ،101 ،89 ،93 89 ،91	0.48
16	Shipp-2002-v1	2	798	418 ،380	0.39
17	Singh-2002	2	339	102 .237	0.38
18	Su-2001	10	1571	98 ،199 ،89 ،101 ،109 ،105، 159 ،162 ،301 ، 248	0.29

The result of our calculations is between the K-Means algorithm with a value of 0.41 and the FMG algorithm with a value of 0.5, which were obtained by averaging the R index, the value of R=0.478, which compared to the outputs of Figure 2.

a)

b)



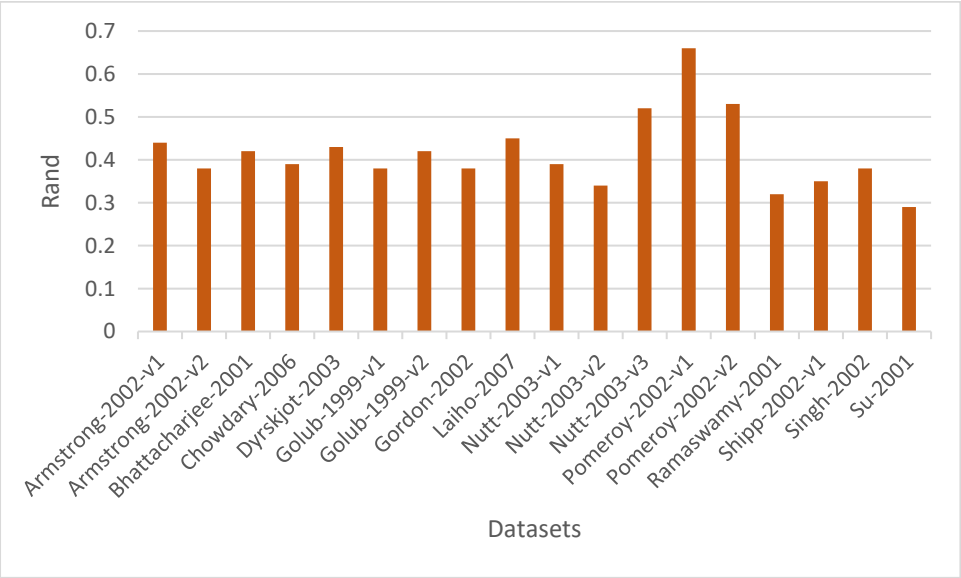


Figure 3: Calculation chart of the R index for each dataset with Euclidean similarity criterion (a) and with a cosine similarity measure (b)

The algorithm was performed again using the cosine similarity criterion (C) for 18 datasets, which displays the result of calculating the Rand index for each of the datasets. The average Rand index was 0.41, which is according to the results of our operations after K_means (Table 4). It should be noted that the output is not available for the FMG algorithm with the cosine similarity criterion (see Figure 3).

Table 4: R index results for each of the datasets by running the algorithm and using the Cosine similarity criterion

No	Dataset	Class number (actual cluster)	Gene number	Gene number in actual clustering	R index
1	Armstrong-2002-v1	2	1081	579،502	0.44
2	Armstrong-2002-v2	3	2194	658 ،805 ،731	0.38
3	Bhattacharjee-2001	5	1543	301 ،128 ،400 ،302 ،401	0.42
4	Chowdary-2006	2	182	77 ،105	0.39
5	Dyrskjot-2003	3	1203	191 ،611 ،401	0.43
6	Golub-1999-v1	2	1877	976 ،901	0.38
7	Golub-1999-v2	3	1877	554 ،815 ،508	0.42
8	Gordon-2002	2	1626	831 ،795	0.38
9	Laiho-2007	2	2202	1308 ،894	0.45
10	Nutt-2003-v1	4	1377	438 ،311 ،297 ،331	0.39
11	Nutt-2003-v2	2	1070	559،511	0.34
12	Nutt-2003-v3	2	1152	646 ،506	0.52
13	Pomeroy-2002-v1	2	857	459 ،398	0.66
14	Pomeroy-2002-v2	5	1379	459 ،314 ،189 ،219 ،205	0.53
15	Ramaswamy-2001	14	1363	،89 ،87 ،111 ،92 ،78 ،101 ،89 ،93 89 ،91 ،69 ،79 ،99	0.32
16	Shipp-2002-v1	2	798	418 ،380	0.35
17	Singh-2002	2	339	102 ،237	0.38

18	Su-2001	10	1571	98 ،199 ،89 ،101 ،109 ،105، 159 ، 162 ،301 ،248	0.29

Conclusion

The use of data balancing techniques is one of the proposed solutions to continue operations. Therefore, the number of records related to each type of cancer is close to each other. It is also possible to use classification technics or methods of combining information, such as Majority voting to improve the results obtained in classifications.

It was shown that the K-means clustering algorithm and its integration with the map-reduce technique are effective in achieving the minimum difference between the actual number of classes in the dataset and the number of recovered clusters. The standard Euclidean Similarity (EZ0) and Cosine Similarity (C) criteria were successfully implemented in this work. It was shown that big data techniques can be effective in detection of types of unknown cancers.

Acknowledgment

This project would not have been complete without the expert guidance of Dr. Amin Irandoost from the computer department of Faculty of Engineering of Islamic Azad University, Hamedan branch for providing the components for experimental implementation and testing. I hereby confrim that, to the best of my knowledge, this work does not pose any conflic of interest to any other research work

References

1. Haskell, C.M., J.S. Berek, and C.M. Haskell, *Cancer treatment*. 1980: Saunders Philadelphia.
2. Wang, Lidong, and Cheryl Ann Alexander. "Big data analytics in healthcare systems." *International Journal of Mathematical, Engineering and Management Sciences* 4.1 (2019): 17.
3. Priyadarshinee, Sudipta, and Madhumita Panda. "Big Data: A Boon to Fight Against Cancer Using Map Reduce Framework." *Smart and Sustainable Technologies: Rural and Tribal Development Using IoT and Cloud Computing*. Springer, Singapore, 2022. 3-8.
4. Shailaja, K., B. Seetharamulu, and M. A. Jabbar. "Prediction of breast cancer using big data analytics." *Engineering & Technology* 7.46 (2018).
5. Rout, Ranjita, and Priyadarsan Parida. "A novel method for melanocytic skin lesion extraction and analysis." *Journal of Discrete Mathematical Sciences and Cryptography* 23.2 (2020): 461-473.
6. Sahoo, Akshya Kumar, and Priyadarsan Parida. "Automatic clustering based approach for brain tumor extraction." *Journal of Physics: Conference Series*. Vol. 1921. No. 1. IOP Publishing, 2021.
7. Sahoo, Akshya Kumar, and Priyadarsan Parida. "A Clustering Based Approach For Meningioma Tumors Extraction From Brain MRI Images." *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*. IEEE, 2020.
8. Nishadi, Thanuja. "AS: Healthcare big data analysis using hadoop MapReduce." *Int. J. Sci. Res. Pub* 9.3 (2019): 87104.
9. Sengupta, Shampa, et al. "Development of a Rice Plant Disease Classification Model in Big Data Environment." *Bioengineering* 9.12 (2022): 758.
10. Gothai, E., et al. "Map-Reduce based Distance Weighted k-Nearest Neighbor Machine Learning Algorithm for Big Data Applications." *Scalable Computing: Practice and Experience* 23.4 (2022): 129-145.
11. Haddad, Omar, Fethi Fkih, and Mohamed Nazih Omri. "Toward a prediction approach based on deep learning in Big Data analytics." *Neural Computing and Applications* (2022): 1-21.
12. Mathew, Juby, and R. Vijaya Kumar. "Multilinear principal component analysis with SVM for disease diagnosis on big data." *IETE Journal of Research* 68.1 (2022): 526-540.
13. Bhagat R, Kumar SS, Shilpa V, Premalata CS, Pallavi VR, Krishnamoorthy L. Aberrant promoter methylation and gene expression of H-cadherin gene is associated with tumor progression and recurrence in epithelial ovarian carcinoma. Clinical Cancer Investigation Journal. 2014 Jul 1;3(4):281.
14. Alshammari FD. Do Non-Viral Microorganisms Play a Role in the Aetiology of Human Cancers? A Review. International Journal of Pharmaceutical Research & Allied Sciences. 2018 Oct 1;7(4).
15. Shahbazian H, Marrefi MS, Arvandi S, Shahbazian N. Investigating the prevalence of anemia and its relation with disease stage and patients' age with cervical cancer referred to Department of Radiotherapy and Oncology of Ahvaz Golestan hospital during 2004-2008. Int J Pharm Res Allied Sci. 2016 Jan 1;5(2):190-3.