Quality Based Analysis of Clustering Algorithms using Diabetes Data for the Prediction of Disease

K. Saravananathan, T. Velmurugan

Abstract*: Clustering is the popular fundamental investigative performance analysis technique commonly used in various applications. The majority of the clustering techniques proved their effectiveness in finding lot of solutions for a variety of datasets. With the aim of test its performance and its clustering qualities are easy to implement by partition based clustering algorithms. The clustering algorithms k-Means and k-Medoids are used to analyze the diabetic datasets and to predict the diseases in this research work. Around 15000 diabetic patient's consequential final bio-chemistry prescription are taken for the diabetes identification. With number of times executed the run time of the algorithms are compared from the different clusters. Based on their performance the first-rate algorithm in each class was found out.. The best suitable algorithm is suggested for the prediction of diabetes data in this work.

Index Terms: Cluster analysis, k-Means clustering, Diabetes Data analysis, , k-Medoids clustering.

I. INTRODUCTION

Medical dataset contains huge quantity of reports regarding prescription and clinical assessment. The medical progression progress repeatedly and all the medical procedures are inexpressible with no taking into consideration time. Hence, time is very significant matter to several clinical field problems. For the research, the k-Means and k-Medoids algorithms are applied in medical database.

The reason for using clustering algorithms on the selected database is aimed to analyze the characteristics of the particular group. The clustering technique is highly helpful to identify the exact solution for maximum number of diabetes patients are affected from diabetic diseases. These two algorithms are going to analyze in the future work for the same dataset.

From this comparison study, it is identified that the best algorithm between mentioned two algorithms. The diabetic dataset is used for the purpose of the comparison of those algorithms. The organization of this paper is described given below: The researcher's related works are discussed in section 2 as a literature survey. The procedure and applications of mentioned algorithms are explained in section 3. The results of the algorithms are mentioned in section 4. At last in section 5 is described the conclusion.

Revised Manuscript Received on August 05, 2019.

K. Saravananathan, SRM Arts and Science College, Kattankulathur, Chennai Email: akilsaro@gmail.com

T. Velmurugan, PG and Research Department of Computer Science and Applications, D.G.Vaishnav College, Chennai

II. LITERATURE SURVEY

Many researchers have been developed a huge number of clustering algorithms and utilized for different useful area in data mining. Many related works have been done on different clinical databases together with diabetic data.

S. Deelers and S. Auwatanamongkol proposed enhancing k-Means algorithm is one of the best algorithms comparing to random initialization in the number of test cases and lesser compilation time of k-Means considerably for massive datasets [1]. P. Padmaja et.al, presented that the taken databases shows the capable shaping from the group of clusters by the number of clustering techniques [2]. T. Velmurugan et.al, concluded that, for smaller datasets k-Means is capable and for large datasets k-Medoids used [3].

T. Velmurugan et.al, mentioned their research, the FCM is better than k-Medoids. The compilation time of two algorithms will increase while the clusters are increases [4]. T. Velmurugan concluded that the k-Means clustering is relatively improved than the Fuzzy C-Means [5]. Mahendra Tiwari et.al, found k-Means algorithm is less execution time for noisy data and k-Medoids is taken number of time for redundancy data [6].

M. Kothainayaki et.al, proposed that the main catchy characteristics of the k-Means are very effective in clustering for big datasets [7]. Abhishek Patel and Purnima Singh found that the change of possible results of the two algorithms based on the process of choosing the first set of mediods [8]. A. Sheshasayee and P. Sharmila presented their research that the more number of data added the execution duration also increased in FCM and k-Means techniques. Therefore, the FCM is the best method comparing to the k-Means algorithm [9].

Dr. T. Karthikeyan and K. Vembandadsamy presented that the hybrid approaches are observed to produce significant results in terms of the classification accuracy, processing time and etc [10]. R. Nithya et al., found that the new outcome it is incidental that the k-Means produces superior performance while comparing other algorithms by using the database [11]. Preeti Arora et al., proposed that the comparison results confirm that time occupied in cluster head selection and space density of overlapping of cluster is much better in k-Medoids than k-Means. [12].

Abdullah M. et al., concluded that the standard techniques in the methods of calculating accuracy and execution time for both the artificial and actual clinical databases [13].

Usha G Biradar et.al, concluded that, in the diabetic dataset gives the best result by using k-Means algorithm [14].

Divya Sharma, et al., found that data mining techniques to help healthcare area, data mining techniques helpful to predict the disease and also provide the exact treatment for the patients [15]. Kalpitet et al., finally concluded their research work, comparing two algorithms the k-Medoids gives excellent result in the execution time [16].

III. MATERIALS AND METHODS

The frequently applied clustering algorithms like k-Means and k-Medoids are considered for the further analysis. The number of inputs and various cluster ranges are utilized in this research. The formations of clusters are maintained certain distance between data points and from its mid points. The group of clusters are differentiated by various colours.

A. Data Set

15,000 patient's medical situation and reports of the diabetes database is getting from the private lab. The dataset used to contain 11 columns of attributes associated to medical diagnosis of a diabetic disease and it mentions the patient is troubled with the diabetes or not. The dataset is divided three dissimilar clusters. The selected two algorithms k-Means and k-Medoids are most suitable to find the exact disease for the diabetic dataset. The attributes of medical diabetes datasets are as follows:

Table I. Input Table

Sl. No.	Attributes
1.	Patient Name
2.	Age
3.	Gender
4.	FPG
5.	PPG
6.	Urea
7.	Creatinine
8.	Sodium
9.	Potassium
10.	HBA1C
11.	Result

B. The Methods

The **k-Means clustering** is a technique utilized to finding clusters in a number of unlikeness or dissimilarity point is reduced. The number of vectors X_j, j = 1,..., n, are to be divide into c groups G_i, i = 1,..., c. The Euclidean distance between a vector X_k in group j and the consequent cluster center c_i, can be described by:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left[\sum_{k,\, X_k \in G_i} \left\| X_k - c_i \right\|^2 \right] \quad (1)$$

The divided groups are noted by the order of (c x n) matrix "U", where the element

$$u_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ data point } X_j \in i \\ 0 & \text{otherwise.} \end{cases}$$

Let c_i be the center of the cluster, the u_{ij} becomes from equation (1) can be derivative as follows:

$$u_{ij} = \begin{cases} 1 & \text{if } \left\| X_j - c_i \right\|^2 \le \left\| X_j - c_k \right\|^2, \text{ for every } k \neq i, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$$c_i = \frac{1}{|G_i|} \sum_{k,\, X_k \in G_i} X_k \quad (3)$$

Where $|G_i| = \sum_{j=1}^{n} u_{ij}$.

A data set X_i, for every value of i varies from 1 to n are applied in the algorithm; consider c_i is the middle of the number of clusters.

k-Medoids Clustering

By the help of PAMs approach, the k clusters are identified from each group of cluster object. When the medoids selected one time, the remaining unselected group of medoids are grouped itself.

The in source of k-Medoids algorithm is containing the number of vectors so the output result of the k-clusters grouped equally selected and others are in separate group of the database. This algorithm comes from the type of alternative translation techniques.

To compute the unrelated group of k-Medoids described below:

Step 1: Take k first points.
Step 2: Assume the outcome of displacing individual of the chosen objects through one of the unselected objects.
Step 3: Choose the pattern among the lowest cost.
Step 4: Otherwise, correlate each unselected point with its nearby chosen point and stop.

The k-Medoids algorithm is based on the search for k representative objects or medoids among the observations of the dataset. With finding a set of k-Medoids, k number of clusters is constructed by assigning each observation to the nearest medoid. The most important plan is to find k-representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object.

IV. EXPERIMENTAL RESULTS

For this study, the frequently used clustering algorithms k-Means and k-Medoids are utilized and observed in diabetes database. By using the mentioned algorithms, the selected 15,000 data are first time divided into five sets of cluster centers. The least and highest values k-Means algorithm is 2949 and 3066, also the least and highest data point of k-Medoids algorithm is 2920 and 3126. These two algorithms utilized ten iterations with the compilation time of k-Means is 931 ms and k-Medoids is 2105 ms.

The compilation time, least and highest data points for the two algorithms are shown in Table II.

Table II. Cluster points

Algorithms	Clusters					Run Time (MS)
	1	2	3	4	5	
k-Means	2988	2994	3066	3003	2949	931
k-Medoids	2920	2975	2967	3012	3126	2105

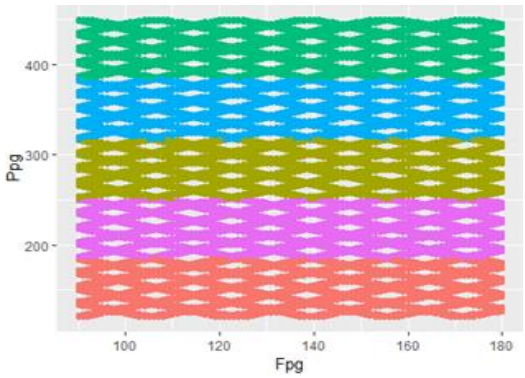

Fig. 1. Clusters of k-Means

From those dataset first analyze age between 20 and 45, Fasting Plasma Glucose (FPG) and Postprandial Plasma Glucose (PPG) ranges are from 90 to 140 and 141 to 450 respectively. The scattered clusters are shown in figures 1 and 2 by the use of two algorithms. Figure 3 shows the comparative study of cluster ranges. From this presentation easily understand the variation of the clusters. From figure 4 knows the time variation.

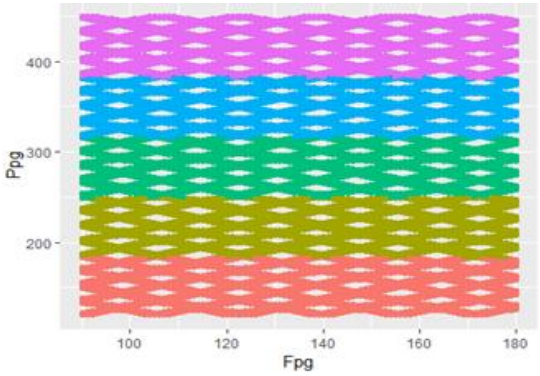

Fig. 2. Clusters by k-Medoids

Fig. 3. Clusters comparison

In next time the dataset is divided into 8 cluster centers by the help of two techniques. From these 8 clusters provides the least value of k-Means is 1,556 and the highest is 2,687.and also the least value of the k-Medoids is 1,521 and the highest value is 2,494.

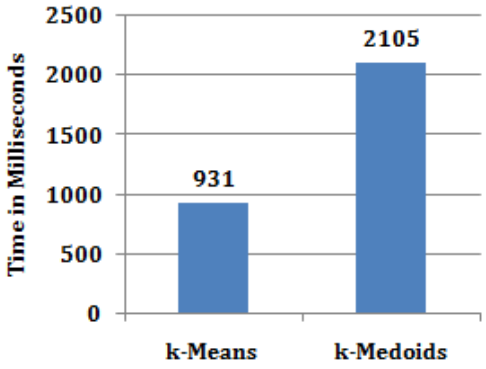

Fig. 4. Time comparison

These two algorithms utilized twelve iterations with the compilation time of k-Means is 1247 ms and k-Medoids is 2288 ms. Table III points that the least and most clusters for the planned two algorithms with compilation time. Figures 5 and 6 mentioned 8 different clusters.

Table III. Cluster ranges

Algorithms	Clusters								Run Time MS
	1	2	3	4	5	6	7	8	
k-Means	1556	1739	1824	1675	1848	1781	1890	2687	1247
k-Medoids	2494	1776	1933	1850	1788	1828	1521	1810	2288

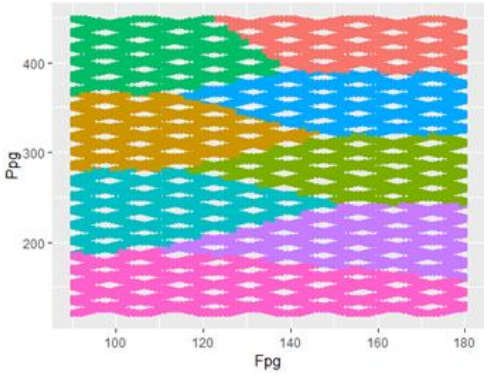

Fig. 5. Clusters in K-Means

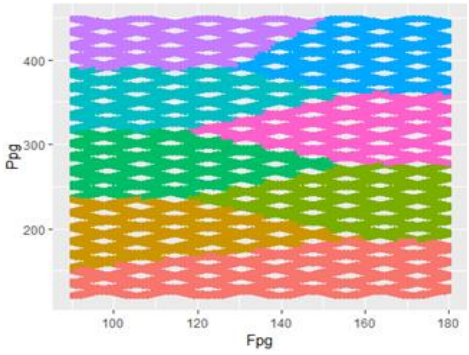

Fig. 6. Clusters in k-Medoids

In this time to compute the age from 30 to 55, the FPG count between 100 and 140 and PPG is 150 to 500. Figure 7 shows the comparison chart. By using the two algorithms the execution time 1247 ms and 2288 ms. Figure 8 described the time comparison.

Fig. 7. Cluster points

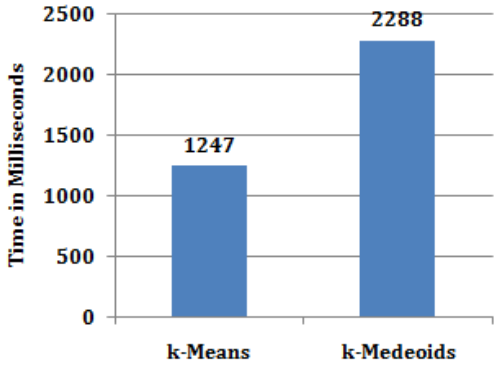

Fig. 8. Time comparison

Here, the datasets are grouped in 10 data clusters by the mentioned above two techniques. From k-Means get the least value is 1,288 and the greatest value is 1,622. In k-Medoids the minimum is 1,326 and maximum is 1608. This time the clustering process is by utilized datasets constraints are the age from 15 to 65, FPG counting range is starting from 90 to 140 and PPG is 141 to 400. The dataset is utilized 15 iterations and calculated the compilation time of k-Means algorithm is 1114 ms and k-Medoids algorithm is 2508 ms. Table IV shows the least and highest values for the planned two algorithms with compilation time. From the figures 9 and 10 are showed the scattered clusters by the help of two algorithms.

Table IV: Cluster ranges

Algorithms	Clusters										Time Ms
	1	2	3	4	5	6	7	8	9	10	
k-Means	1622	1614	1435	1590	1507	1532	1489	1288	1535	1388	1114
k-Medoids	1607	1326	1417	1521	1537	1592	1553	1608	1353	1486	2503

Figure 11 shows the plotter for the clusters. By using these two algorithms the total execution time is 1114 ms and 2503 ms, and the figure 12 shows the time comparison chart.

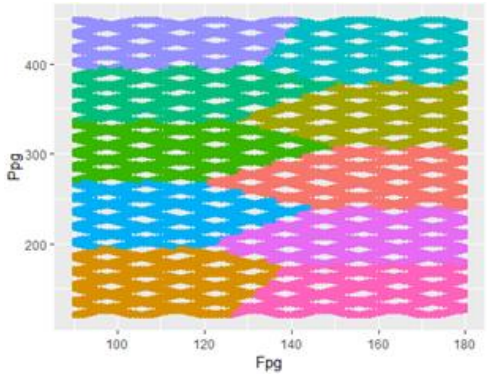

Fig. 9. Clustering using k-Means

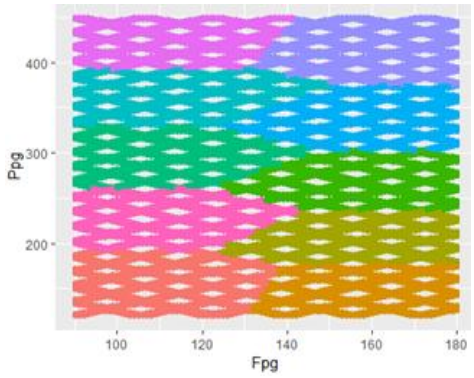

Fig. 10. Clustering using k-Medoids

Fig. 11. Results of cluster points

Table V shows the execution time of two techniques. Figure 13 shows the time comparison between two algorithms with three different cluster points.

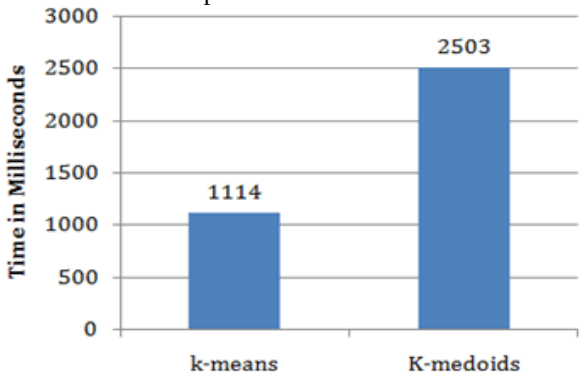

Fig. 12. Comparison Chart

Table V. Time Comparison of two algorithms with three different clusters

Algorithms	Number of Clusters		
	5 Clusters	8 Clusters	10 Clusters
k-Means	931 ms	1247 ms	1114 ms
k-Medoids	2105 ms	2288 ms	2503 ms

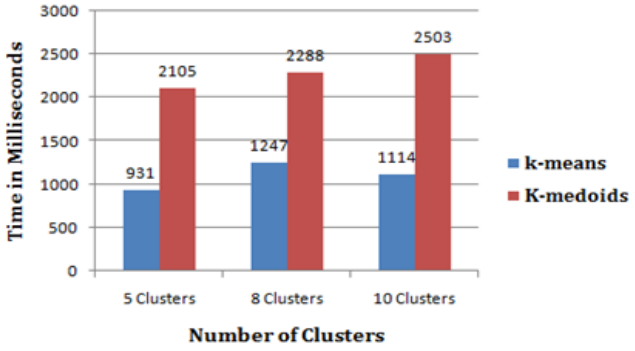

Fig. 13. Comparison of two algorithms in terms of different clusters

The accuracy of two algorithms is calculated based on the following measures. From those dataset to analyze FPG ranges between 90 and 140. PPG ranges between 140 and 450. k-Means accuracy is 87% and k-Mediods accuracy is 80 %, so that k-Means algorithm is one of the best techniques compared with k-Medoids.

V. CONCLUSIONS

The comparative study of the compilation time and accuracy of the k-Means and k-Medoids clustering algorithms are obtained. By comparing the execution time and accuracy using of diabetic dataset are utilized for the analysis in order to experiment of the k-Means and k-Medoids clustering algorithms. For this research, patient's past and before food blood glucose diagnosis report data clustered and analyzed. The physicians, lab testers, and clinical experts are easily predicting the diabetic disease with help of this analysis. For these clustering algorithms the diabetic patient's report data are given as input attributes. In this analysis approach, the disease of the patients is absolutely identified based on their medical report and it is confirmed with physicians and medical experts. This research mentioned k-Means clustering is better than the k-Mediods clustering as a result of their compilation time and accuracy. By this reason, the k-Means clustering is better than k-Medoids clustering. From the same dataset, the upcoming research is to improve the performance of other algorithms.

REFERENCES

1. S. Deelers, and S. Auwatanamongkol, "Enhancing k-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance", International Journal of Physical and Mathematical Sciences, Vol. 1 (11), 2007, pp. 518 – 523.
2. P.Padmaja, Srikanth Vikkurty, NiloferInaz Siddiqui, Praveen Dasari, Bikkina Ambica, V.B.V.E. Venkata Rao, Mastan Vali Shaik, and V.J.P. Raju Rudraraju, "Characteristic Evaluation Of Diabetes Data Using Clustering Techniques**",** International Journal of Computer Science and Network Security, Vol.8 (11),2008, pp. 244 – 251.
3. T. Velmurugan and T. Santhanam, "Computational Complexity between k-Means and k-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points", Journal of Computer Science, Vol. 6 (3), 2010, pp. 363 – 368.
4. T. Velmurugan and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach", Information Technology Journal, Vol. 10 (3), 2011, pp. 478 – 484.
5. Dr. T. Velmurugan, "Performance Comparison between k-Means and Fuzzy C-Means Algorithms using Arbitrary Data Points", Wulfenia Journal, Vol. 19 (8), 2012, pp. 234 – 241.
6. Mahendra Tiwari and Randhir Singh, "Comparative Investigation of k-Means and k-Medoid Algorithm on Iris Data", International Journal of Engineering Research and Development, Vol. 4 (8), 2012, pp. 69 – 72.
7. M. Kothainayaki and P. Thangaraj, "Clustering and Classifying Diabetic Data Sets Using k-Means Algorithm", Journal of Applied Information Science, Vol. 1 (1), 2013, pp. 23 – 27.
8. Abhishek Patel and Purnima Singh, "New Approach for k-Means and k-Medoids Algorithm", International Journal of Computer Applications Technology and Research, Vol. 2 (1), 2013, pp. 1 – 5.
9. Sheshasayee A, Sharmila P., "Comparative study of Fuzzy C-Means and k-Means algorithm for requirements clustering", Indian Journal of Science and Technology, Vol. 7 (6), 2014, pp. 853–857.
10. Dr. T. Karthikeyan and K. Vembandadsamy, "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus", International Journal of Computer Application, Vol. 5 (5), 2015, pp. 96 – 104.
11. R. Nithya, P. Manikandan, Dr. D. Ramyachitra, "Analysis of clustering technique for the diabetes dataset using the training set parameter", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4 (9), 2015, pp. 166 – 169.
12. Preeti Arora, Dr. Deepali, and Shipra Varshney, "Analysis of k-Means and k-Medoids Algorithm for Big Data", International Conference on Information Security & Privacy, 2015, pp. 507 – 512.
13. Abdullah M. Iliyasu, Chastine Fatichah, Khaled A. Abuhasel, "Evidence Accumulation Clustering with Possibilitic Fuzzy C-Means base clustering approach to disease diagnosis", Automatika, Vol. 57 (1), 2016, pp. 822 – 835.
14. Usha G Biradar and Deepa S Mugali, "Clustering Algorithms on Diabetes Data: Comparative Case Study", International Journal of Advanced Research in Computer Science, Vol. 8 (5), 2017, pp, 550-552.
15. Divya Sharma, Anand Sharma, and Vibhakar Mansotra, "A Literature Survey on Data Mining Techniques to Predict Lifestyle Diseases", International Journal for Research in Applied Science & Engineering Technology, Vol. 5 (4), 2017, pp. 1575 – 1581.
16. Kalpit, G. Soni, and Dr. Atul Patel, "Comparative Analysis of k-Means and k-Medoids Algorithm on IRIS Data", International Journal of Computational Intelligence Research, Vol. 13 (5), 2017, pp. 899-906.

AUTHORS PROFILE

K. Saravananathan is working as an Assistant Professor in the Department of Computer Science and Applications, SRM Arts and Science College, Kattankulathur, Chennai – 603 203, Tami Nadu, India. Now he is doing Ph.D. degree in Computer Science in the University of Madras. He has published three papers in different Indian and International Journals. Also, he hss participated workshops and presented papers in various conferences.

Dr. T. Velmurugan is working as an Associate Professor in the PG and Research Department of Computer Science and Applications, D.G.Vaishnav College, Chennai – 600 106, India. Also, he is the Head of the Department of Computer Science and Applications (UG). He holds a Ph.D. degree in Computer Science from the University of Madras. He elected as a Senate Member, University of Madras. He has published more than 90 articles indexed in SCOPUS and SCI such as Applied Soft Computing, Journal of Computer Science and etc. He was an invited speaker of the 10th Int'l Conference on Computational Intelligence and Software Engineering (CiSE 2018) held from January 5-7, 2018 in Bangkok, Thailand. He has guided more than 300 M.Phil. Research Scholars in the field of Computer Science. He guided 7 Ph.D. scholars and currently guiding 10 Ph.D. scholars in the same field. He served as a nominated Senate Member in Middle East University, Dubai, UAE for a period of three years. He is a member in Board of studies for many autonomous institutions and Universities like Periyar University, Salem, India. He hosted a lot of programs in Doordharsan television about recent topics in Information Technology field. He arranged and acted as an Organizing Secretary of International Conference on Computing and Intelligence Systems (ICCIS 2015). In addition, he was a resource person for various national workshops entitled "Scientific Research Article Writing and Journal Publications". He is an editorial board member of many International Journals. He is also a reviewer in many peer reviewed journals like Elsevier, Springer, IOS Press Journals etc. Further, he is a visiting faculty for M.Phil. course for various universities throughout India.