*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 20, Issue 3, Ver. III (May. - June. 2018), PP 61-67*
*www.iosrjournals.org*

# Ranked Cluster Based Adaptive Sampling with Gradient Boosting Classifier for Medical Data

B. Senthil Kumar [1], Dr. R. Gunavathi [2]
[1] *Assistant Professor, Department of Computer Science, Sree Narayana Guru College* [2] *Associate Professor & Head, Department of MCA, Sree Saraswathi Thyagaraja College*
*Corresponding Auther: B. Senthil Kumar*

***Abstract:*** *Classification is the process of supervised learning process, which is an important task in pattern recognition. Imbalanced class distribution is always a trouble in achieving high accurate classification. The imbalanced data is more complicated for the medical data classification. There are several algorithms and techniques have been used to overcome the class imbalance issues. However, the techniques can be either insufficient nor leads to misclassification for medical datasets. To overcome the issue of class imbalance in the medical data classification, an effective adaptive sampling technique is proposed. The technique is named as Ranked Cluster Based Adaptive Sampling with Gradient Boosting classifier (RCBAS-GBC). The technique is quite adequate for medical datasets. The experiments are carried with several unbalanced datasets and achieved good results.*

***Index Terms:*** *Classification, class imbalance, sampling Techniques, Medical data mining.*

---

Date of Submission: 09-06-2018 Date of acceptance: 25-06-2018

---

## I. Introduction

Diabetic disease is the most frequently diagnosed disease in India. The clinic care unit utilizes several data mining techniques and tools for diabetic disease oriented decision making [1]. For the appropriate decision making, different training samples are used. But the major drawback of such approaches and techniques are training sample inequality, which crates class imbalance issues in the classification. This proposal aims to overcome the issues of class imbalance in disease classification [2]. Class imbalance is a major issue in the classification due to the imbalance dataset for each class. Most of the Machine learning algorithms are optimal when there will be equal number of instances for each class. If the count is mismatch, then it creates a class imbalance issue. Handling class imbalance issues in medical data classification using sampling techniques is the main aim of the proposed system.

There are different types of class imbalance issue solving techniques are proposed. The techniques are known as under sampling and oversampling techniques [3]. Oversampling techniques adds training samples in the minority class, whereas under sampling eliminates the excess training samples from the majority classes. However, the techniques are not fully adopted for all type of data classifications. This creates over generalization problems, where leads to misclassification issues. The existing approach to handle imbalanced data is depicted in fig 1.0.

Other widely used techniques are based on sampling. They are used for selecting a representative subset of relevant data from a large dataset. In many cases, sampling is very useful because processing the entire dataset is computationally too expensive. In general, the critical issue of these strategies is the selection of a limited but representative sample from the entire dataset. Various random, deterministic, density biased sampling, pseudo-random number generator and sampling from non-uniform distribution strategies exist in the literature [4]. However, very little work has been done on the Pseudorandom number generator and sampling from nonuniform distribution strategies, especially in the multi-dimensional case with heterogeneous data. Naive sampling methods are not suitable for noisy data which are part of real-world applications, since the performance of the algorithms may vary unpredictably and significantly. The random sampling approach effectively ignores all the information present in the samples which are not part of the reduced subset [5]. An advanced data reduction algorithm should be developed in multi-dimensional real-world datasets, taking into account the heterogeneous aspect of the data. Both approaches [6] [7] are based on sampling and a probabilistic representation from uniform distribution strategies.

The authors of [8] proposed a method to reduce the complexity of solving Partially Observable Markov Decision Processes (POMDP) in continuous state spaces. The paper uses sampling techniques to reduce the complexity of the POMDPs by reducing the number of state variables on the basis of samples drawn from these distributions by means of a Monte Carlo approach and conditional distributions.

The authors in [9] applied dimensionality reduction to a recent movement representation used in robotics, called Probabilistic Movement Primitives (ProMP), and they addressed the problem of fitting a low-dimensional, probabilistic representation to a set of demonstrations of a task.

**Existing Approach to handling Imbalanced Datasets**

| Techniques | Drawbacks |
|---|---|
| **Data Level approach: Re-sampling Techniques** | It increases the likelihood of over fitting since it replicates the minority class events. |
| **Random Over-Sampling** | It replicates the minority class events. |
| **Informed Over Sampling: Synthetic Minority Over-sampling Technique** | Increase in overlapping of classes and can introduce additional noise.<br>Not very effective for high dimensional data |
| **Modified synthetic minority oversampling technique (MSMOTE** | Noise over data |
| **Cluster-Based Over Sampling** | Possibility of over-fitting the training data |

The authors fitted the trajectory distributions and estimated the parameters with a model-based stochastic using the maximum likelihood method. This method assumes that the data follow a multivariate normal distribution which is different from the typical assumptions about the relationship between the empirical data. The best we can do is to examine the sensitivity of results for different assumptions about the data distribution and estimate the optimal space dimension of the data.

## II. Proposed Work

The proposed system provides a sampling method called **Ranked Cluster Based Adaptive Sampling with Gradient Boosting classifier (RCBAS-GBC)**, to minimize the workload of diabetes disease classification by identifying appropriate training over sampling. The objective is to handle the class imbalance problem that occurs due to varying size of different disease dataset. It detects the appropriate percentage for sampling process to improve the classification accuracy. The proposed method finds improved performance in terms of performance metrics called F-measure and Area under Curve (AUC).

**Proposed approach**

Combines the proposed sampling technique with the gradient boosting classifier

**Ranked cluster adaptive sampling technique**

**Gradient Boosting classifier**

The proposed system overcomes the above drawbacks by developing a new effective over sampling technique **Ranked Cluster Based Adaptive Sampling with Gradient Boosting classifier (RCBAS-GBC)**, which is falls under the **cluster based oversampling** technique.

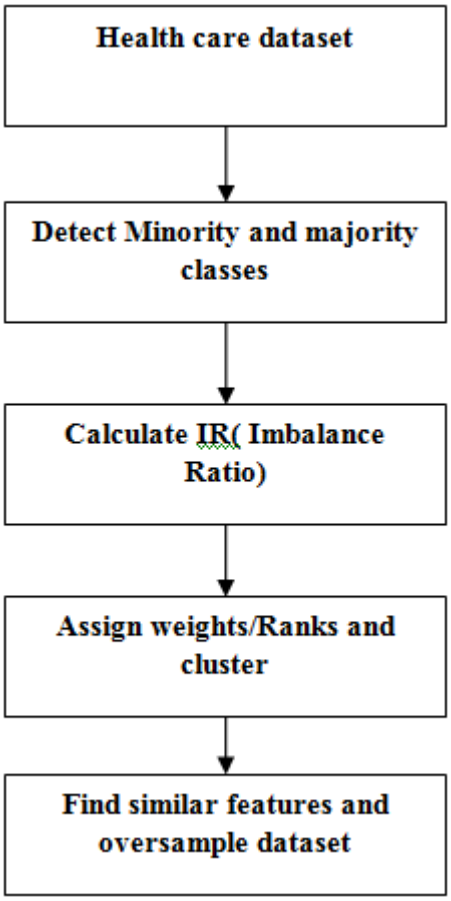


**Figure 3.0 steps of RCBAS**

**2.1 RCBAS:**

**RCBAS** technique was used to analyze the informative minority and majority set. The following steps present the flow of the process.

1. Initially the filtered minority set D may be identified from the original minority set D . To try this, Similarity Index SI(x ), for each data sample x ∈ D was computed. After that, each x will be eliminated if its Similarity Index SI(x ) carries only the majority class samples. The eliminated minority class sample known as noisy data where it is entirely surrounded by the majority class samples. From this it is sure that RCBAS removes noises effectively and also prevent the process from the noisy data.

For each data sample x ∈ D , RCBAS will build a nearest majority set calledN (x ). The samples in N (x ) can be the borderline majorities and expected to be placed near the decision boundary while the nearest majority k2 samples is small. Then all the N (x ) was combined in order that it forms about the borderline majority set, D .

For every value of y ∈ D RCBAS constructs N (y ) and combines all such N (y ) to form D. The parameter k3 utilized in N (y ) needs to be effectively adequate for including minority synthetic class samples required to generate by using RCBAS sampling approach. The disease classification using RCBAS technique is given in figure 3.0: The medical data is considered in which the selection of minority class samples is being done by using RCBAS Technique.

**2.1.1 WEIGHT VALUE ASSIGNMENT**

The class imbalance issue is handled by the RCBAS shows the construction of a set of the informative minority and majority class samples, Cs to be carried out for the generation of the adaptive data samples. Nonetheless, all the samples of this elegance may not be as considerable. A few samples may also offer extra beneficial information to the facts than the others. Therefore, it is important for assigning weights to the samples in keeping with their significance. A large weight value implies that the synthetic samples generated from and

nearby of informative samples is due to the insufficiency of information in its minority concept. It is now understood that Cs is to be computed by thinking about the predefined observations. The RCBAS considers them and uses the majority class set Cs in evaluating Cs. Each majority class sample Ai ∈ Csmaj provides a weight to every minority class sample B ∈ Cs and vice versa. This weight is known as the information weight Iw (Ai, Bi). For each value of B all the values of Iw (Ai, Bi) were summed up to find its selection weight known as Cs (Bi).

**Notations:**

l =the dimension of the feature space.
d=distance function
D=density function
P=Proximity

This can be expressed in equation 1.0.

$$Cs(Bi) = \sum Ai \epsilon Csmaj\ Iw(Ai, Bi) \quad (1.0)$$

While considering the process of RCBAS, I (A , B ) is calculated as the product of the proximity factor, P (A , B ) and the density factor, D (A , B ) given by the formula,

$$I(A, B) = P(A, B) \times D(A, B) \quad (2.0)$$

The minority class samples having more majority class neighbors in Cs will receive a higher selection weight value. The adaptation property allows the modification based on density factor. The following process demonstrates how the RCBAS computes the factor.

The Proximity factor: P (A, B ) The computation of the proximity factor is very straight forward.

Steps:

1. Normalized Euclidean distance D (A, B) =dist (Ai,Bi)/l was initially computed.

Where dist (A , B ) is the Euclidean distance measured from A to B .

2. P (A, B) value is computed in the following way:

$$P(A, B) = \frac{t\left(\frac{1}{dn\ (Ai, Bi)}\right) + Csmax}{P(t)} \quad (3.0)$$

Where t and Csmax are the user defined parameters and t is the cut-off function. The inverse of the normalized Euclidean distance is first applied to f in this equation. This is performed to ignore the values that are too high and for slicing them to the highest value defined by $\left(\frac{1}{dn\ (Ai, Bi)}\right)$. The value that was found would lie between the ranges [0, csmax]. It is defined in the following equation 4.0.

$$t(B) = \{B \text{ if } B \leq t \text{ otherwise } t \quad (4.0)$$

The Density Factor: D (A, B) The density factor implies that the sparse cluster should possess many number of synthetic samples than the dense cluster. It is known that both the clusters are equally balanced from the decision boundary, and it adopts based on the density value. Hence, RCBAS computes the value of D (A, B) by normalizing the value of t (A, B) given in equation 5.0 below.

$$D(A, B) = \frac{t(A,B)}{Eq\,\epsilon Csmin\ (t(A,B))} \quad (5.0)$$

The existing methodology called graph cut clustering algorithm cannot handle these problems in case of dynamic growth of data. The graph cut clustering (GCC) technique has been previously utilized for the clustering of medical data however the problem of processing entire graph every time even when the graph undergoes minor changes makes it unsuitable. Hence a modified rank clustering (RC) method is introduced.

### 2.1.2 RANKED CLUSTERING APPROACH

The healthcare dataset largely dependent on the attribute and its samples, in the existing research work, many classifications and as well as clustering algorithms were used for partitioning the graph data in an efficient manner. In case of dynamic data updated in the training samples, the existing work lack of efficient partitioning. It was quiet difficult and time consuming for the record updation in the dataset. Concisely saying the existing system does not perform well in the case of dynamic growth of medical data. and the dataset updation and oversampling process need a ranked dataset to be updated in the training samples.

Hence a Ranked Clustering (RC) is introduced that encourages the insertion and deletion of newly arrived data into the graph in the efficient manner. RC used an increment clustering set of rules which matches with the aid of placing the 2 variables over each vertex (clusters). These variables are, "in cluster weight and out cluster weight". In cluster weight is determined as the overall of weights of edges which might be joining the vertices that are identified within the same bunch. The variable out cluster weight is determined as the total weights of edges that are interlinked with the vertices present in one-of-a-kind clusters. The incremental clustering can assist 4 varieties of operations. They are Edge Insertions, Edge Deletion, Vertex Insertion, and Vertex Deletion. Edge clustering will be smooth process while it is far delivered for the connection of vertices found in equal cluster. For this type of insertion adjacency matrix replace is enough. However if the edge is added that is interconnecting the vertices present in among the clusters, then it's far vital to replace the clusters.

The cluster obtained after RCBAS approach will decide latest records of data samples in different classes. Each and every cluster will consist of equivalent data samples in every class. This cluster information may be used to perform the classification task.

**2.2 GBC ALGORITHM**

GBC is based on the combination of cluster-based adaptive sampling and gradient boosting algorithm. It is similar to RUSBoost and SMOTEBoost with the critical difference occurring in the sampling technique and the iterations used. SMOTEBoost uses SMOTE method to oversample the minority class instances, while RUSBoost uses random under-sampling on the majority class. In comparison, the proposed GBC uses cluster-based ranked adaptive sampling from the class imbalance issue. GBC finds the class inequality and performs adaptive sampling process by using the **RCBAS** algorithm. Here, the threshold is determined by **RCBAS**. After that, ranking process will be performed and sampling process will be selected based on the inequality ratio. As clustering and ranking is used before sampling and finally the oversampled datasets are used in the classification.

## III. Experimental Results

The medical data sets is considered in this work consists of attributes like different types of disease dataset for eg., diabetes, heart disease, liver diseases. Such data are collected from the UCI repositories. By using these datasets the different types of disease dataset are analyzed by the same algorithm to find the effectiveness of the proposed work. The Collected data are imbalanced, in order to make such data as balanced data by applying Adaptive Over sampling technique in the data. In this work RCBAS with the existing algorithm called SMOTE were compared. The performance evaluation is done based on the performance metrics such as oversampling processing time, and accuracy. This performance analysis is represented in the graphical representation. The over-sampled data will be used in the classification process. The initial step of the implementation process is applying the RC approach and the outcome of this clustering will give the classification process.

The experiments were conducted using three different datasets such as Pima Indian diabetic, which has been modified to perform class imbalance issue and heart dataset. The dataset consist of 2 classes and different attributes collected from UCI repository and modified to evaluate the proposed method. The sampling process and classification process accuracy is greater than 90% shown in Table 1.0.

**Table 1.0 accuracy for different iterations**

| Iterations | Accuracy |
|---|---|
| 1 | 92.8 |
| 2 | 95.2 |
| 3 | 98.6 |

The proposed adaptive sampling technique is measured with the accuracy factor with various iterations. Here the iteration defines the number of attempts made for effective sampling process and classification. The result shows the different threshold with optimal data will increase the accuracy of the proposed system. The proposed system has maximum three iterations.

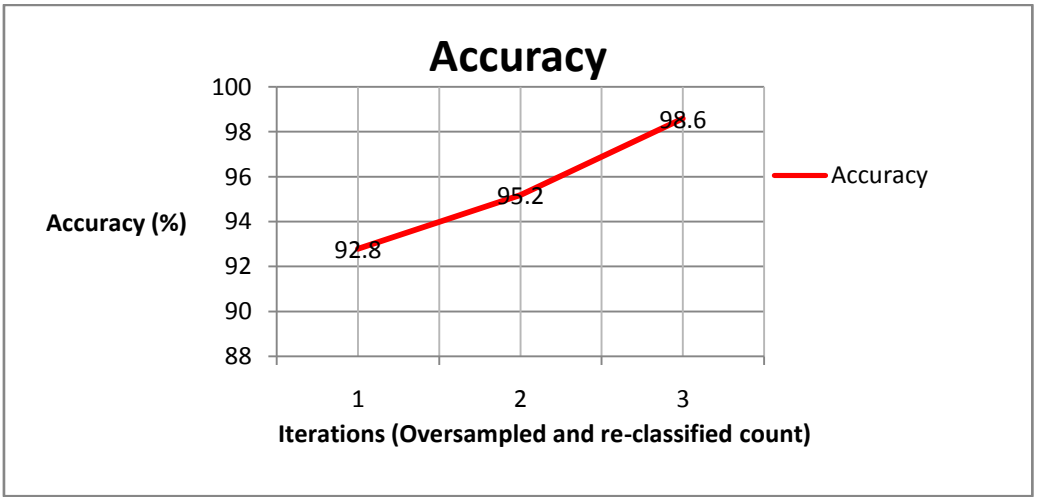


**Figure 4.0 Iterations and accuracy**

An Intel dual core 2 duo with 3.0 GHz processor with 2 Gb RAM was used to measure the execution time and memory usage. Table 2.0 tabulates the execution time for varying data imbalance count and Table 3.0 gives the memory usage for varying imbalanced datasets.

**Table 2.0 Execution time for varying imbalance data samples**

| Imbalance ratio Count | Execution time in second |
|---|---|
| 80 | 21.8 |
| 70 | 18.2 |
| 60 | 14.6 |
| 50 | 13.5 |
| 40 | 10.7 |
| 30 | 8.1 |
| 20 | 7.4 |
| 10 | 6 |
| 5 | 3.9 |

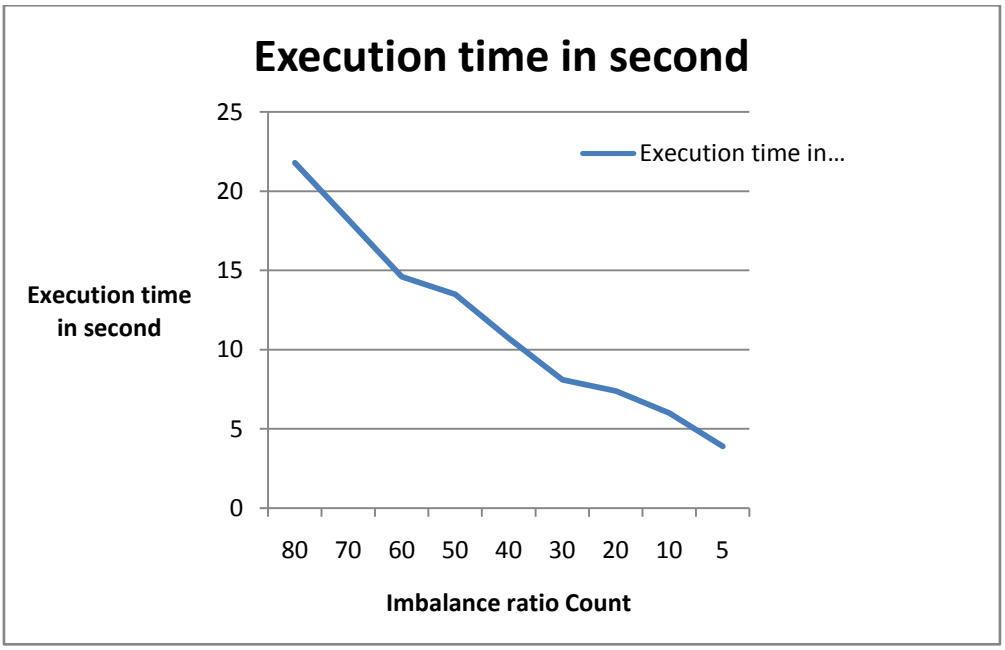


**Figure 5.0 Execution time vs imbalance ratio**

**Table 3.0 Memory usage for varying imbalance data sets**

| Imbalance ratio Count | Memory usage (Mb) |
|---|---|
| 80 | 746 |
| 70 | 719 |
| 60 | 719 |
| 50 | 636 |
| 40 | 567 |
| 30 | 359 |
| 20 | 193 |
| 10 | 152 |
| 5 | 83 |

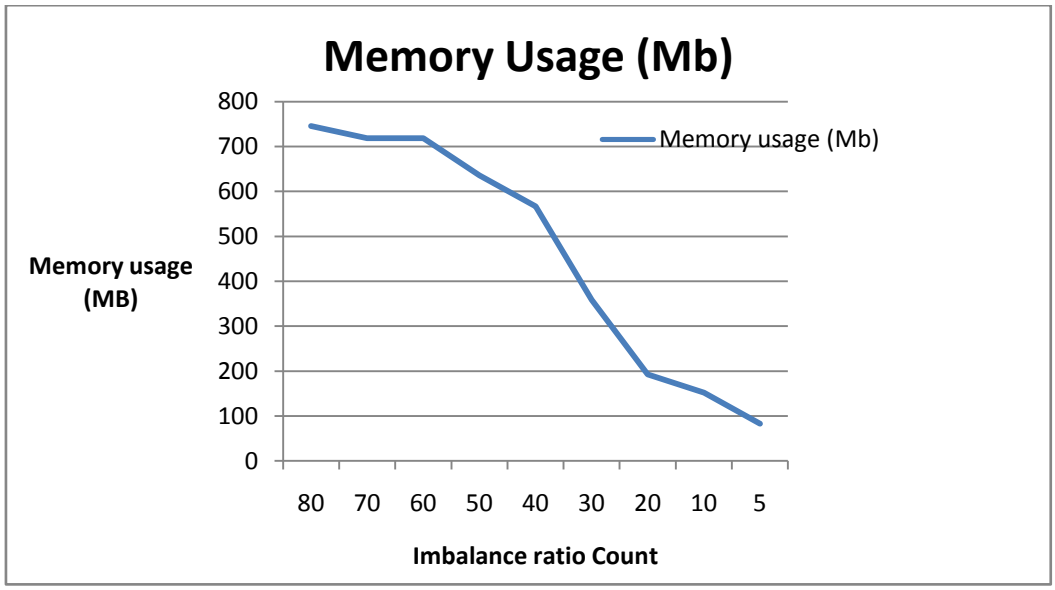


**Figure 3.8 Memory usage and imbalance ratio count**

It was proposed to find imbalance and adaptive ratio to tackle the imbalance dataset among different classes. Associations based on the class ratio is calculated and adapted. A novel **Ranked Cluster Based Adaptive Sampling** (RCBAS) is proposed and classification has been performed on the modified dataset. The proposed method at accuracy levels greater than 93% was able to identify the exact class even in case of imbalanced data's. Different types of comparisons were carried out to evaluate the performance with respect to imbalance data ratio, execution time and memory usage. It is evaluated for different types of medical datasets. It was observed that the imbalance ratio decrease the execution time and memory usage reduces.

## IV. Conclusion

In this paper, the ranked cluster based adaptive sampling is presented to reduce the work classification especially in medical dataset. It is done by introducing a methodology called the **Ranked Cluster Based Adaptive Sampling** (RCBAS) algorithm for adaptive sampling, which performs both under and over sampling based on the dataset. The experimental tests were conducted and analyzed and compared with the existing methodology called SMOTE method. The performance analysis is made by comparing it with the existing methodology and it is proved that the proposed method improves in its performance over metrics such as accuracy, execution time and memory.

## References

[1]. B. Senthil, and R. Gunavathi."A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis." IJARCCE. 5. 463-467. 10.17148/IJARCCE.2016.512105 (2016).
[2]. Kumar, B. Senthil, and R. Gunavathi. "Comparative and Analysis of Classification Problems." Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 7.8 (2017).
[3]. Chawla, Nitesh V. "Data mining for imbalanced datasets: An overview." Data mining and knowledge discovery handbook. Springer, Boston, MA, 2009. 875-886.
[4]. Rubinstein, R. Y., & Kroese, D. P. (2011). Simulation and the Monte Carlo method volume 707. John Wiley & Sons
[5]. Whelan, M., Khac, N. A. L., Kechadi, M. et al. (2010). Data reduction in very large spatio-temporal datasets. In Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on (pp. 104–109). IEEE.
[6]. Mejía-Lavalle, Manuel, et al. "Toward Optimal Pedagogical Action Patterns by Means of Partially Observable Markov Decision Process." Mexican International Conference on Artificial Intelligence. Springer, Cham, 2016.
[7]. McGhan, Catharine LR, Ali Nasir, and Ella M. Atkins. "Human intent prediction using markov decision processes." Journal of Aerospace Information Systems (2015).
[8]. Fakoor, R., & Huber, M. (2012). A sampling-based approach to reduce the complexity of continuous state space POMDPs by decomposition into coupled perceptual and decision processes. In Machine Learning and Applications (ICMLA), 2012 11th International Conference on (pp. 687–692). IEEE volume 1.
[9]. Colom´e, A., Neumann, G., Peters, J., & Torras, C. (2014). Dimensionality reduction for probabilistic movement primitives. In Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on (pp. 794– 800). IEEE.

B. Senthil Kumar "Ranked Cluster Based Adaptive Sampling with Gradient Boosting Classifier for Medical Data "IOSR Journal of Computer Engineering (IOSR-JCE) 20.3 (2018): 61-67.