Journal of Green Engineering (JGE)

Volume-10, Issue-12, December 2020

Disease Detection using Enhanced K-Means Clustering and Davies-Bould in Index in Big Data –Safe Engineering Application

[1*] G. Uday Kiran and [2]D. Vasumathi

[1]*Assistant Professor, Department of Computer Science and Engineering, B V Raju Institute of Technology, Vishnupur, Narsapur, Telangana, India.*
E-mail: udaykiran.goru@bvrit.ac.in
[2]*Professor, Department of Computer Science and Engineering, JNTU College of Engineering, Hyderabad, India. E-mail: rochan44@gmail.com*

Abstract

Data clustering is a significant technique used in distinct fields for measuring the similarity present among the data which is useful in day to day life applications. Clustering is the process where the dataset will be divided into number of groups of same data points. However, in few instances it would undergo the problem of overlapping because many features may not be ready to capture necessary information in order to separate clusters. In order to overcome the limitations, the Enhanced K-Means Clustering (EKMC) method is proposed for data clustering. To perform classification on highly imbalanced datasets, the proposed EKMC algorithm is used by computing the distance metrics between the two different data points. Therefore, in the proposed method, an improvised KMC known as an Enhanced KMC overcomes the problem occurred in KMC for disease detection. The KMC has the specialty of producing acceptable results however, it fails to provide good results as the data consists of outliers the spread density data points across the space is different. This differences shows reduction in CPU and memory requirements. K-means clustering algorithm can be significantly improved by using a better initialization technique such as Davies–Bouldin Index (DBI). DBI computes the ratio between the cluster distances and the between cluster distances thereby calculates average of overall clusters. The EKMC performs various iterations sequentially and in each iteration KMC computes the distances among the data points and the centers which consume more time, also expensive due to their huge UCI repository datasets. In order

Journal of Green Engineering, Vol. 10_12, 13089-13106.

to overcome the problem, the EKMC is computed in the next iteration which uses the values obtained in the previous nearest cluster based on the distance. This results are evaluated and the proposed method achieves accuracy of 96.71% when compared with the existing KMC method.

Keywords: Data clustering, Davies–Bouldin Index, Enhanced K-Means Clustering, Imbalanced, Overlapping.

1 Introduction

Prediction of healthcare has increased the accurate diagnosing and also allows to maintaining good public healthiness. By the help of big data concept, many researchers have developed the prediction models, however they have too many cases present in the dataset, it is bit difficult to analyze in lesser time. Many researchers have used national and international databases for examining and identifying the big data analytics for predicting the different sort of diseases such as heart attack, lung cancer, breast cancer etc., [1-2]. The large amount of datasets is used to perform medical treatment and also diagnosed the process which is a significant field in data mining and pattern recognition. Medical datasets used in the existing researches for automatic identification has shown improvement in the system accuracy, the reliability, and speed of the system [3-4]. The results obtained from the machine learning and artificial intelligence techniques shows the full potential and also an effective preparation for performing an experiment. By using the redundant features in the existing methods a decreased accuracy was generated. Therefore, a proper selection of features or auto selected features techniques need to be developed. Thus, along with data preparation, a proper classification method is need to be achieve high accuracy in disease prediction [5-7]. The problems such as processing large training data is performed by using a novel EKMC method thereby reduces the training data set that selects which is useful in subset data reduction technology [8]. The EKMC method used in the research that modifies K-means clustering which operated based on distinct methods by diagnose the disease. The result evaluates various performance measures obtained that shows whether the type of disease occurred is of malignant or benign [9]. K-means clustering algorithm significantly improves a better initialization technique such as Davies–Bouldin Index (DBI). An evaluation scheme is validated which will tell how well the quantities have been clustered and how much extent the features are inherited from the dataset as per the proposed method.

The structure of the paper is as follows: Section 2 is the Literature Review that describes about the existing methods undergone for distinct disease prediction. Section 3 explains about the proposed EKMC-DBI method with the help of block diagram. Section 4 explains about the results and discussion for the proposed EKMC-DBI and also a comparative study is evaluated. Finally, section 5 is the conclusion of the research and the future work needed to be done further.

2 Literature Review

Recently, various researches have experience for disease detection using KMC with improvised models. They are as follows:

Nan et al. [10] developed a clustering algorithm known as Chinese Herbal Medicine, with the optimization algorithm Artificial Bee Colony (ABC). The Chinese Herbal Medicine methods were integrated with ABC that gave rise to a novel strategy for searching the neighbor nectar (please go through the cited paper). The developed model optimized the parameters employed from the developed improvised ABC algorithm. The experiments used in UCI and TCM datasets shows an effective result in terms of accuracy when compared with the traditional classical clustering algorithms (k-means, k-Mediods). However, the developed model could not meet better effort in targeting for larger TCM datasets using ABC.

Zhu et al. [11] developed an improvised logistic regression model for prediction of diabetes that integrated Principal Component Analysis (PCA) and K-means techniques. The results obtained by using PCA showed enhancement by using the k-means clustering algorithm and logistic regression was used for classification in terms of accuracy with a k-means output classified the data correctly. However, the results achieved were compared with various researchers that showed that the developed method limited the data availability for final prediction and classification.

Lin et al. [12] developed a Reversible privacy-preserving clustering technique based on k -means algorithm. The developed Privacy-Preserving Data Mining (PPDM) model that contain the original data and is used by various researchers. The developed PPDM used methods like swapping, modifications and deletion for the data that were protected by the original data thereby protected the resultant data. However, the developed k-means clustering with PPDM as Reversible Privacy Preserving (RPP) shows better results in terms of performance measures but when corresponded to grouping clusters results were at risk which reduced the performance values.

Guruler [13] developed a diagnosing system in order to detect Parkinson's disease. The complex valued Artificial Neural Network (ANN) with K-Means Clustering Based Feature Weighting (KMCFW) method. The KMCFW combined with Complex Valued ANN (CVANN) generated features. These newly obtained features are converted into a number of complex values. The features are having values that were fed as an input for CVANN. The results showed that the effectiveness and efficiency of the developed system was evaluated enormously against the Parkinson's disease dataset that significantly achieved the highest classification results. However, the developed system used the simple-valued (please go through the cited paper) classifiers which did not give a positive impression in terms of providing an exact and quick diagnosis of PD.

Le et al. [14] developed a K-means Interval Type-2 Fuzzy Neural Network (IT2FNN) algorithm for medical diagnosis. Firstly, the test data were fed for the classifier for classification used for determining the best

classification. The developed IT2FNN obtained the parameters with the help of steepest decent approach. The Lyapunov theory was used in the developed method for convergence and stability verification. However, the developed model was not applicable for all the classification processed systems and also, the computation process was tougher and was also hard to implement.

3 Proposed Method

The block diagram of the proposed method is shown the representation of the medical diagnosis. The Medical diagnosis determines the disease or condition of a patients based on their symptoms. Data sets include a vast amount of medical data, various measurements, financial data and etc. (This is general description). The main sources of medical statistics are medical records. The block diagram of the proposed research is shown in the figure 1.

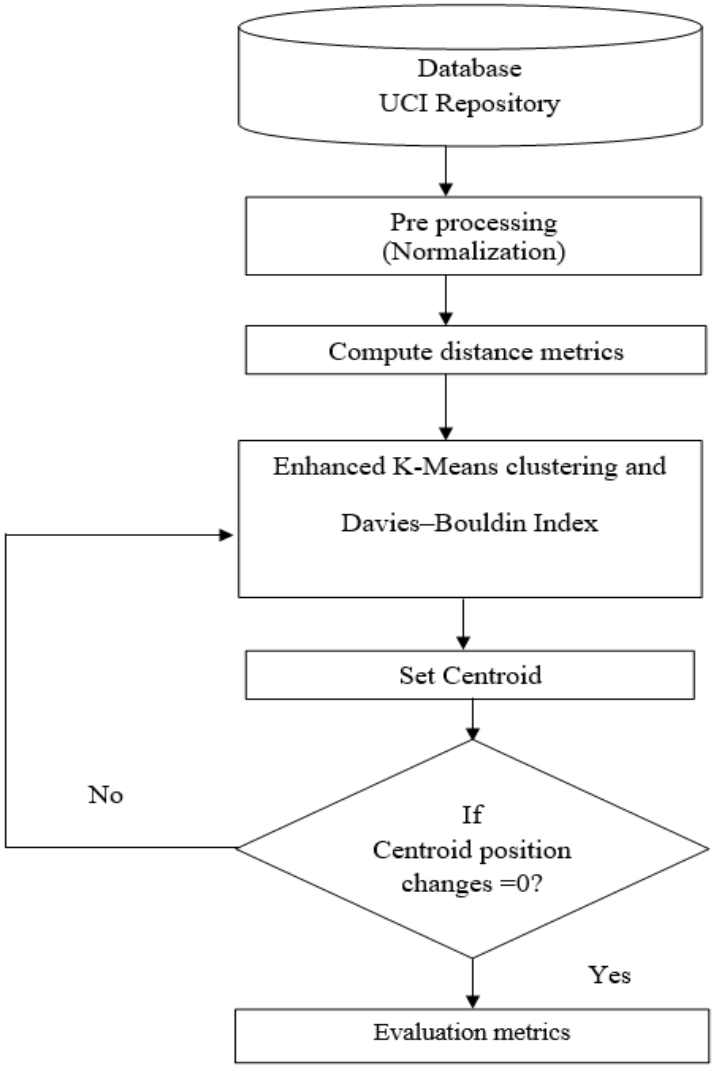

Figure 1 Block diagram of the proposed EKMC-DBI method

3.1 Dataset

The UCI machine learning Repository database consisted of collection of domain theories; database and data generators are used in the machine learning approaches that analyses empirical analysis for different algorithms [15]. The dimension range of UCI dataset starts from 6 to 61 that include continuous data but also discrete and continuous mixed data. The data were collected from a Level-three hospital in Heilongjiang Province of China in 2006-2012, that includes 955 records from patients. These records are collected from the medical database that consists of medical report which are used to process for preprocessing step. The databases are collected for several medical diseases from UCI data repository like liver disorder, breast cancer, Indian liver patients, heart disease, hepatitis, Parkinson's, dermatology and lung cancer. The pre-process is performed by using the normalization process in order to compensate the problem of missing data or human errors during data acquisition.

3.2 Data Preprocessing

The preprocessing collects the raw data from the UCI repository database and converts into the understandable medical report. The examination of large number of patient's data is considered for the proposed research in order to get better accuracy values when more number of training data is provided. Usually, the data consists of missing data in large amount because of human error. In order to avoid this, the missing data will be filled by the structured data. The incomplete or uncertain medical data is present and should be modified by deleting and to improve the quality of the data. The Min-Max normalization process plays an important role for integration and as well as data normalization. Each and every feature value has minimum value which gets transformed into 0 and the maximum value is transformed into 1. All the values will be converted into decimals that ranges from 0 to 1. The equation for the normalization process is as follows.

$$X_{norm} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \qquad (1)$$

Where, X_i is the ith data point, X_{min} is the minimum value of the data point, $X_{max}$is the maximum value of the data point. These variables calculate normalized value known as X_{norm} which are used for filling the missing data uniformly forms a structured data.

Once the min max normalization is performed for the unstructured data, the uncertainty in the data will be present. Thus, extraction of such features from the various complex structures helps to determine predication of disease.

3.3 Enhanced K-means Clustering

The K-means algorithm is the unsupervised algorithm which is used to classify the clusters. These clusters are known as k-clusters which are used to minimize the squared Error for performing optimal classification. The k-means algorithm has an advantage of improving the system speed, robustness and simplicity. The K-means algorithm is efficient when compared to other clustering algorithms which perform well when compared to linear dataset. It is comparatively simple to implement and faster than hierarchical clustering. The K-means algorithm minimizes the sum of distances from every object to its cluster centroid. This algorithm runs until the sum of distance cannot be decreased further.

Step 1: Start
Step 2: Select the number of cluster center
Step 3: Set initial cluster center randomly
Step 4: Put object to the closest cluster center
Step 5: Recalculate the new cluster center
Step 6: Create cluster based on smallest distance
Step 7: If the objects move to clusters show an output else return to step 2 to initialize cluster center randomly.

(Detailed explanation is only given below know)

Step 1: Choose the initial cluster centers 'k' among all the clusters is defined as that are randomly selected by k samples with a given set of sample.

$$Z_1(1), Z_2(2), ..., Z_k(k) \tag{2}$$

Step 2: By distributing the samples {x} by k^{th} iterative phase among the K cluster domains that uses the following relation :

$$x \in S_j(k) \tag{3}$$

if

$$\left\| x - Z_j(k) \right\| < \left\| x - Z_i(k) \right\| \tag{4}$$

Where all $i = 1,2,3,\ldots..,\ K$
$i \neq j$,
$S_j(k)$ is defined as the set of sampled clusters by $Z_j(k)$

Step 3: By Computing the new clustering center defined as, $Z_j(k), j = 1,2,.....k$ the total sum of all the new clusters that were centered is reduced. The new cluster center) $Z_j(k+1)$ is computed using the performance index by the minimization which is given as follows:

$$J_j = \sum_{x \in S_j(k)} \left\| x - Z_j(k+1) \right\|^2 \tag{5}$$

Where $j = 1, 2, ...K$

The performance index (minimum) is the mean of $S_j(k)$ that gives the cluster center as a new is as follows:

$$Z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} \tag{6}$$

Where $i = 1, 2, 3, ...K$

N_j is the number of $S_j(k)$ and thus the k-means derivate from the updated cluster modes were sequenced.

Step 4: If the equation (7) executes then the output is obtained and once after the output obtained, the procedure is terminated. Else if the condition is not satisfied step 2 is going to be executed.

$$Z_j(k+1) = Z_j(k) \tag{7}$$

Where

$j = 1, 2, ...K$

This KMC behaviour is understood based on the functionality of specified cluster centers. The initial cluster centers made choice that makes a sample order of clusters which define their geometrical properties. The KMC has the specialty of producing acceptable results. However, it fails to provide good results as the data consists of outliers that spread density data points across the space. This shows difference in reduction for CPU time and memory requirements for a large dataset. In order to overcome the problem, an improvised KMC is implemented known EKMC which overcomes the problem occurred in existing KMC for disease detection. The EKMC algorithm requires simple data structure which programmed to store the data and perform various operations. It is easy to implement for each iteration when operated and the results obtained from the previous iterations are used for the next iterations. The results illustrated that EKMC improved the computational speed, Magnitude of total distance calculations when compared with existing KMC. The role of centroid is that it avoids the overlap of two data points are occurred

- If the centroid position is not changed, the metrics will be evaluated.
- If the centroid position is changed, initialization of EKMC will be executed.

In the section, an idea of the proposed method which makes k-means significant for large consisting of large number of clusters. For each iteration, the EKMC computes the distances between the data points and the centers which consume more time, and also expensive (problem) due to their huge datasets. In the upcoming iteration in order to overcome the problem happened in the first iteration, distance is computed for previous nearest cluster. This checks again for two conditions

If new distance is <= the previous distance
(Point stays in their cluster) **# Execute**
Else
Not necessary to compute the distance for other cluster centers**# Execute**

By executing the aforementioned conditions, the time required to compute distances of k−1 cluster center will be saved.

In the proposed method, we mentioned or classify into two functions such as first and second function. They are as follows:

First Function

The basic function of k-means algorithm is used to determine the centre for nearest data points thereby computes the distances of k as centers. In each data point, the distance will be kept at the nearest centre. A simple data structure keeps the distance between each data point and their cluster center. This distance is calculated by using Euclidian distance () given by the equation (8) and (9)

$$d(i,j)=d(j,i)=\sqrt{(i_1-j_1)^2+(i_2-j_2)^2+...+(i_n-j_n)^2} \quad (8)$$

$$=\sqrt{\sum_{n=1}^{k}(i_k-j_k)^2} \quad (9)$$

Where $d(\ )$ is the distance parameter; i and j are the cluster

The steps followed will be as follows:

- The distance between the data points i and k centroids.
- The function finds the closest centroid for the point number i and are calculated for the closest centroid j as well
- The data points i is added to the j number clusters
- The count of the points are incremented as $j+1$
- The proposed idea is used to analyze and the number of closest cluster present are used for calculating the new distance which is represented as recalculation distance_new ()

Once the new distance is calculated second function is executed

Pseudo Code for the First Function

Def (first_function):
Compute the distance among the i data points having k as the centroids using Equation (9)
Calculate the closed centroid having point number i
    Calculate for the j closest centroid j
For all Data points i is added to the j clusters:
    Increment ()
    $j+1$
first_function =Calculate $distance\_new()$

Second Function

- The distance among the current data point i and the new cluster centre will be assigned to the previous iteration that used for computing the smaller distance which could be smaller or equal to the older center. The assigned point cluster is stayed in the previous iteration and it is not required to compute the distances to the other clusters having $k-1$ centers.
- If the distance which is computed is huge when compared to the old cluster center, there will be change in the cluster distance among the current points with respect to all k centered clusters.
- In the next step the closest center is searched
- Current point to the closest cluster is assigned to increase the count of points in cluster.
- In order to determine the closest cluster quicker, DBI is determined which is the ratio between the within cluster distances and the between cluster distances and computing the average overall the clusters.

Pseudo Code for the Second Function

Def (second_function):
Calculate the distance among the current data point i and new cluster in second_function is calculated
If the new cluster distance < first_function new_distance:
    Negotiate the new_distance value
Else:
    Calculate the distance for other clusters having $k-1$ centers
Elif:
    Calculate the closest cluster centre using DBI
second_function = Calculate $distance\_new()$

The DBI is defined as the ratio of distances within a cluster and between the clusters that average the sum of all the clusters. Therefore, it is simple for computing the bound between 0 to 1 lower score. When the data receives are well separated by clusters, the performance of k-means are executed. DBI is usually used for evaluating the clustering algorithms that is used for internal evaluation thereby validates clustering that takes the features and quantities inherent from the dataset. DBI is given by an equation (10)

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} \left| X_j - A_i \right| \right)^{\frac{1}{p}} \qquad (10)$$

Where n are the dimensional points.
S_i is the measures the scatter among the cluster.
C_j is the data points of cluster.

X_j is an n dimensional feature vector

A_i is the centroid of C_j

T_i is the cluster size

P = 2 as it is considering Euclidean_distance () function among the cluster centroids for every feature vectors.

The distance metrics are used for higher dimensional data finds the Euclidean distance that determines the better measure for cluster determination. The distance metric matches the metric is used for clustering scheme which gives meaningful results by using the Equations (11) and (12).

$$M_{i,j} = \left\| A_i - A_j \right\|_p \quad (11)$$

$$= \left(\sum_{k=1}^{n} \left| a_{k,i} - a_{k,j} \right|^p \right)^{\frac{1}{p}} \quad (12)$$

Where,

$M_{i,j}$ is the measure that separates among C_i and C_j

$a_{k,i}$ is the element k, A_i for n^{th} dimension centroid, A is the n centroid dimensions.

k Indexes for the data features that determines the Euclidean distance among clusters i and j

Let $R_{i,j}$ will be the measure for good clustering scheme that gives the definition for the account $M_{i,j}$ measure among the clusters i and j.

S_j is cluster that scatter for i^{th} cluster which is low as possible which is given by using the following properties as in equations (13-16)

$$R_{i,j} \geq 0 \quad (13)$$

$$R_{i,j} = R_{j,i} \quad (14)$$

When $S_j \geq S_k$ and $M_{i,j} = M_{i,k}$ then $R_{i,j} > R_{i,k}$ (15)

When $S_j \geq S_k$ and $M_{i,j} \leq M_{i,k}$ then $R_{i,j} > R_{i,k}$ (16)

According to this formulation, the value lowered and it gives a separation among the clusters and tightness is obtained from the clusters. The solution for these properties are defined as shown below equation (17).

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (17)$$

To define, D_i following below equation (18) is used

$$D_i = \max_{j \neq i} R_{i,j} \quad (18)$$

If in case N is the cluster numbers then it is defined as by using the below equation (19)

$$DBI = \frac{1}{N}\sum_{i=1}^{N} D_i \tag{19}$$

The value of DBI is dependent on both EKMC and data which finds the best most similar cluster using the equation (18). The similarity is selected based on cluster and weighted average and so on. The steps followed to reduce the calculation of distance as each point is assigned based on the closest cluster that is used to calculate the faster function that calculates the function distance. In the second step an implementation function that calculates the function distance () which executes two times with continuous iterations used to calculate the distance_new (). This acts as a reminder of iteration with the help of DBI for the overall average clusters. These 2 steps executed is known as "enhanced k-means" algorithm.

4 Results and Discussion

The proposed method uses Python 3.7 tool for the experiment. The experiments are conducted on an Intel Core i7 processor having 2 GHz CPU and 48GB memory. The classification results are computed for the data which is divided into testing and training data. The results obtained for the proposed scheme shows better results in terms of Recall, Precision, Accuracy, F-score for the research.

4.1 Performance Measures

The performance measures, Accuracy, precision, F-measure, Recall are calculated in terms of parameters such as TP is the True Positive, TN is the True Negative, FP is the False Positive, FN is the False Negative values.

- **Precision**

Precision is also known as Positive Predictive Value which is the ratio of relevant instances to the retrieved instances defined by using the equation (20)

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

- **Recall**

Recall is defined as the number Recall of relevant documents retrieved to search divided by the total number of existing relevant documents defined by using the equation (21)

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

- **F1-Score**

F-score are statistical variability that performs representation of random errors defined by using the equation (22)

$$F - Score = \frac{2TP}{\left(2TP + FP + FN\right)} \tag{22}$$

- **Accuracy**

Accuracy is defined as the ratio of accurately predicted to the total number of observations. The accuracy is calculated by using the equation (23)

$$Accuracy = \frac{\left(TP + TN\right)}{\left(TP + TN + FP + FN\right)} \tag{23}$$

4.2 Quantitative Analysis

The proposed EKMC-DBI is proposed in the research in order to perform disease prediction using different types of disease where the datasets are taken from UCI. The distinct types of datasets for disease includes liver disorder, breast cancer, hepatitis, heart disease, Parkinson's, Lung cancer and Dermatology. The dermatology dataset has achieved lowest accuracy of 95.89 % when compared with other 6 types of datasets. Whereas, heart disease has achieved the best accuracy of 97.48 % when compared with other 6 types of datasets. Similarly, Hepatitis has achieved lowest accuracy of 94.21% when compared to other 6 datasets whereas lung cancer has achieved the best accuracy of 96.8% when compared with other 6 datasets. The graphical representation for the table 1 is shown in the figure 2.

Table 1 The results obtained for the proposed Enhanced KMC-DBI in terms of Accuracy and Precision

	UCI Dataset	Accuracy (%)	Precision (%)
Disease	**Liver disorder**	96.36	95.38
	Breast cancer	97.21	96.21
	Heart disease	97.48	96.58
	Hepatitis	95.42	94.21
	Parkinson's	97.25	96.54
	Dermatology	95.89	94.37
	Lung cancer	97.36	96.8

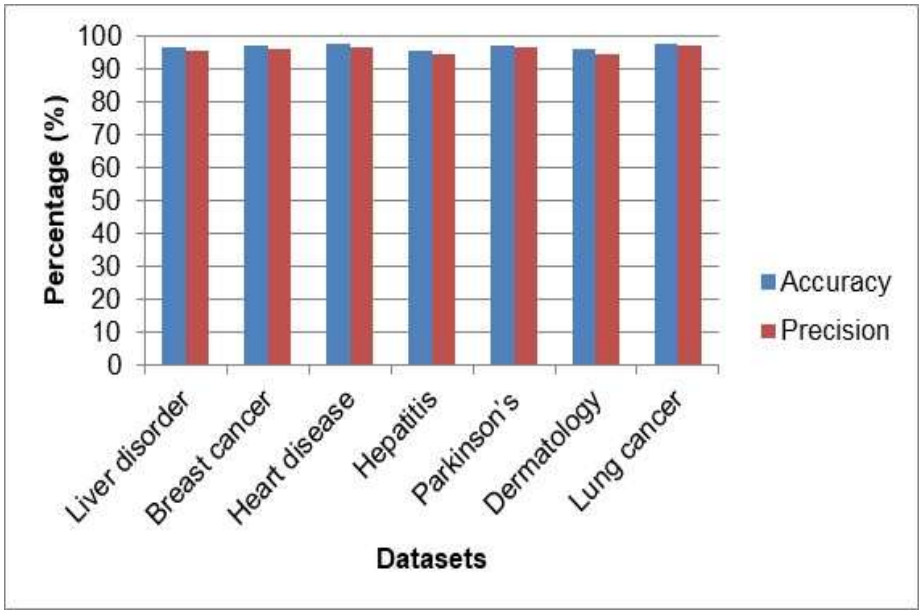

Figure 2 Evaluated values of the proposed EKMC DBI with respect to Accuracy and Precision with a graphical representation

In similar, the results obtained for the 7 datasets is shown in table 2 which are expressed in Recall and F-measure. The dermatology dataset has achieved lowest Recall for Hepatitis of 95.33 % when compared with other 6 types of datasets. Whereas, heart disease has achieved the best Recall for Breast cancer 96.84 when compared with other 6 types of datasets. Similarly, Hepatitis has achieved lowest F- measure of 94.76 % when compared to other 6 datasets. Whereas, heart disease has achieved the best F-measure of 96.64 % when compared with other 6 datasets. The graphical representation for the table 2 is shown in the figure 3.

Table 2 The results obtained for the proposed Enhanced KMC-DBI in terms of Recall and F-measure.

UCI Dataset	Recall (%)	F-measure (%)
Liver disorder	96.23	95.80
Breast cancer	96.84	96.52
Heart disease	96.72	96.64
Hepatitis	95.33	94.76
Parkinson's	95.88	96.20
Dermatology	95.42	94.89
Lung cancer	96.11	96.45

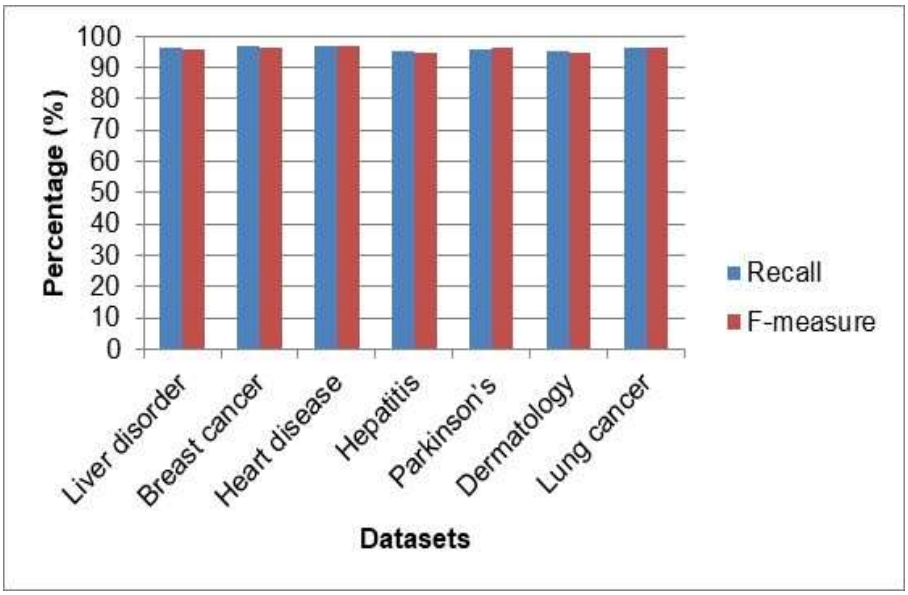

Figure 3 Evaluated values of the proposed EKMC DBI with respect to Recall and F-measure with a graphical representation

5 Comparative Analysis

The table 3 shows the comparative analysis made between the existing and the proposed EKMC-DBI method. The existing method Interval type-2 Fuzzy Neural Network with KMC used for detecting the disease at an early stage based on the disease data. An average accuracy of 93.81 % was achieved by the developed KMC method. Similarly, another existing method Modified K-Means Algorithm with Relief Algorithm was developed that achieved an average accuracy of 90% for disease prediction. But the proposed EKMC method achieved the best accuracy of 96.71 % and also showed improvement when compared with the existing methods. In the existing methods Fuzzy Neural Network with KMC, Modified K-Means Algorithm with Relief Algorithm, the cluster distances were not exactly determined whereas the proposed EKMC DB covered all the clusters distances thereby improved the accuracy and also saved the computation time. The graphical representation of the existing and the proposed EKMC DBI is represented as shown in the figure 4.

Table 3 Comparative results for the proposed and existing methods achieved in terms of Accuracy

Authors	Methodology	Accuracy (%)
Le et al. [14]	Interval type-2 Fuzzy Neural Network with KMC	93.81
Inan et al. [4]	Modified K-Means Algorithm with Relief Algorithm	90.00
Proposed	**Enhanced KMC-DBI**	**96.71**

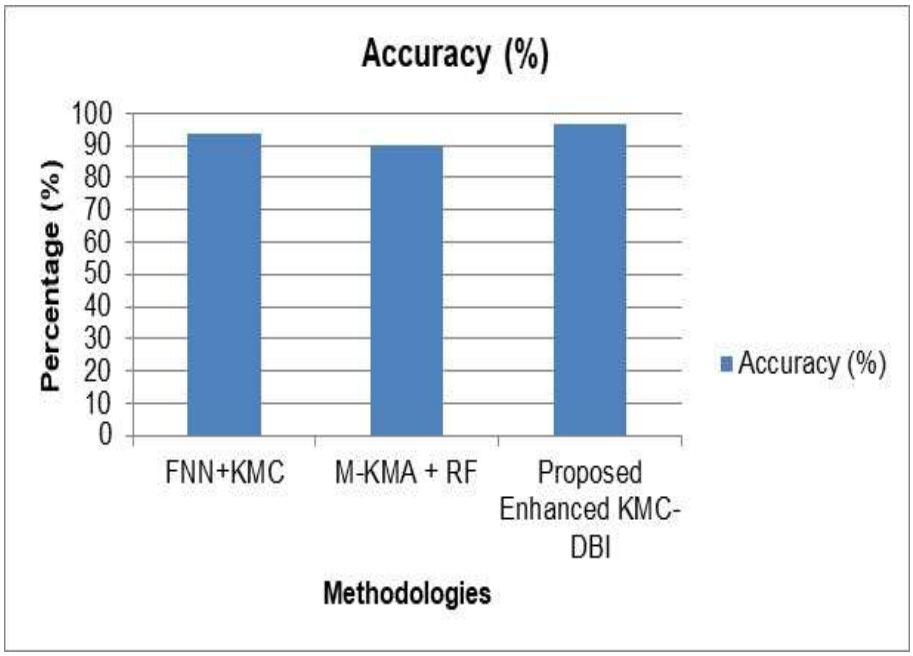

Figure 4 Comparison table for the proposed EKMC DBI and existing methods in terms of Accuracy with a graphical representation

6 Conclusion

Many researchers have used national and international databases for examining and identifying the big data analytics for predicting the different sort of diseases. However, due to overlapping problems during classification of clusters the accuracy was not achieved better. In order to overcome such an issue, the proposed method introduces EKMC method. Initially, the UCI databases are collected for several medical diseases from UCI data repository like liver disorder, breast cancer, Indian liver patients, heart disease, hepatitis, Parkinson's, dermatology and lung cancer. Later, the integration of

data is done by data pre-processing. Min-Max normalization process plays an important role for integration and as well as in data normalization. The KMC has the specialty of producing acceptable results however, it fails to provide good results as the data consists of outliers the spread density data points across the space. This reflected on reduction in CPU and memory requirements for a large dataset. In order to overcome the problem occurred in the existing KMC, t, an improvised KMC is implemented known EKMC is implemented for disease detection. The EKMC algorithm is easy to implement, requiring a simple data structure to keep some information during each iteration execution and the obtained results are used for the next iteration. In order to determine the closest cluster quicker, DBI is determined which is the ratio between the within cluster distances and the between cluster distances and computing the average overall the clusters. The results and discussion for the proposed EKMC-DBI and also a comparative study is valuated. The proposed EKMC method 3% to 6% of accuracy improvement when compared with the existing KMC.

References

[1] R. Venkatesh, C. Balasubramanian, M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine Learning Technique", Journal of medical systems, Vol. 43, no. 8, pp. 272, 2019.

[2] M. J. Sousa, A. M. Pesqueira, C. Lemos, M. Sousa, A. Rocha, "Decision-making based on big data analytics for people management in healthcare organizations", Journal of medical systems, Vol. 43, no. 9, pp. 290, 2019.

[3] V. Tang, P. K. Y. Siu, K. L. Choy, H. Y. Lam, G. T. S. Ho, C. K. M. Lee, Y.P. Tsang, "An adaptive clinical decision support system for serving the elderly with chronic diseases in healthcare industry", Expert Systems, Vol. 36, no. 2, pp. e12369, 2019.

[4] O. Inan, M. S. Uzer, "A Method of Classification Performance Improvement via a Strategy of Clustering-Based Data Elimination Integrated with k-Fold Cross-Validation", Arabian Journal for Science and Engineering, 2020.

[5] M. S. Amin, Y. K. Chiam, K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", Telematics and Informatics, Vol. 36, pp. 82-93, 2020.

[6] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, J.A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}} $ Statistical Model and Optimally Configured Deep Neural Network", IEEE Access, Vol. 7, pp. 34938-34945, 2019.

[7] K. Mittal, G. Aggarwal, P. Mahajan, "Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy", International Journal of Information Technology, Vol. 11, No. 3, pp. 535-540, 2019.

[8] T. Tang, S. Chen, M. Zhao, W. Huang, J. Luo, "Very large-scale data classification based on K-means clustering and multi-kernel SVM", Soft Computing, Vol. 23, No. 11, pp. 3793-3801, 2019.

[9] K. Mittal, G. Aggarwal, P. Mahajan, "Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy", International Journal of Information Technology, Vol. 11, no. 3, pp. 535-540, 2019.

[10] N. Han, S. Qiao, G. Yuan, P. Huang, D. Liu, K. Yue, "A novel Chinese herbal medicine clustering algorithm via artificial bee colony optimization", Artificial Intelligence in Medicine, Vol. 101, pp. 101760, 2019.

[11] C. Zhu, C. U. Idemudia, W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques", Informatics in Medicine Unlocked, Vol. 17, pp. 100179, 2019.

[12] C. Y. Lin, "A reversible privacy-preserving clustering technique based on k-means algorithm", Applied Soft Computing, Vol. 87, pp. 105995, 2020.

[13] H. Gürüler, "A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method", Neural Computing and Applications, Vol. 28, no. 7, pp. 1657-1666, 2019.

[14] T. L. Le, T. T. Huynh, L. Y. Lin, C. M. Lin, F. Chao, "A K-means Interval Type-2 Fuzzy Neural Network for Medical Diagnosis", International Journal of Fuzzy Systems, Vol. 21, no. 7, pp. 2258-2269, 2019.

[15] Newman, D. J., Hettich, S., Blake, C. L. S., & Merz, C. J., "UCI repository of machine learning database", Irvine, CA: University of California, 1988.

Biographies

G. Uday Kiran, completed his B.Tech (CSE) in 2006 from Jawaharlal Nehru Technological University Hyderabad and M.Tech (Neural Networks) in 2010 from Jawaharlal Nehru Technological University Kakinada. He is

working as Assistant Professor in Department of Computer Science and Engineering at B V Raju Institute of Technology, Narsapur, Medak (District). He has 12 Years of teaching experience. His areas of interests are Data Mining, Deep Learning, Computer Vision and Natural Language Processing.

D. Vasumathi, completed her B.Tech, and M.Tech from Jawaharlal Nehru Technological University Hyderabad. She completed Ph.D in Data Mining from Jawaharlal Nehru Technological University Hyderabad. Presently, she is working as Professor in Department of Computer Science and Engineering and she is also Board of Studies Chairman for Department of Computer Science and Engineering at JNTU College of Engineering, Kukatpally, Hyderabad and has more than 20 Years of experience in teaching and experience. She is a member for several professional bodies like CSI, IEEE and ISTE. She had presented and published several papers in National and International Conferences and also in IEEE Explorer. She was chair for several conferences. She did for Editorial board member for several papers of National and International events. Her area of interest includes Data Mining, Big Data and Analytics and Machine Learning.