IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Disease Prediction System using Improved K-means Clustering Algorithm and Machine Learning

C. Kaur[1*], K. Sharma[2], A.K. Sohal [3]

[1] Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India
[2] Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India
[3] Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India

[*]*Corresponding Author: chandanpreet1995@gmail.com, Tel.: +91-88722-82209*

DOI: https://doi.org/10.26438/ijcse/v7i5.11481153 | Available online at: www.ijcseonline.org

Accepted: 24/May/2019, Published: 31/May/2019

Abstract— Now-a-days data mining is widely used in the medical field for analysis and diagnosis of disease. Various techniques such as clustering, classification, association of data mining are used to disclose unseen patterns from large number of datasets. Data mining techniques are applied on incompetent medical data recorded on daily basis. These techniques help to get useful information for diagnosing the diseases. Generally, numbers of tests are required to know the presence of a disease. In order to reduce these numbers of tests, data mining is utilized. In this paper, benign and malignant type of data for breast cancer disease has been used in which Benign tumour is non-cancerous tumour and malignant is cancerous tumour. In this research, two approaches are implemented in MATLAB for disease prediction. The first approach is based on k-means clustering and SVM algorithm for classification algorithm. In second approach, improved k-means clustering algorithm and SVM algorithm is implemented. The second approach gives better performance in terms of accuracy. Accuracy of classification of dataset depends upon the optimization of clustering and pre-processing of dataset.

Keywords— Data Mining, K-means, SVM

I. INTRODUCTION

Data mining concerns with the discovery of information from large amount of raw data. It deals with the exposure of unseen patterns, knowledge and new rules from large set of databases. It is used to convert the unstructured data into structured data, so that knowledge from this data can be used to get unknown facts and rules. Data mining is a key component of knowledge discovery in databases which is closely related to data warehouse. Data warehouse is a huge repository of raw data extracted from different operational systems. Single database referred as data warehouse has integrated large amount of data from multiple sources. Data warehouse is subject oriented, non-volatile, integrated and time variant. Data mining extracts knowledge from data warehouse. Various steps under data mining include data collection, data cleaning, data selection and data analysis. Data collection is to collect or gather data from various sources includes internet, organization, people etc. Data cleaning is to refine the collected data so that noisy and unwanted data can be removed. Data selection includes dimensionality reduction which has two methods: data selection and data extraction. Usually datasets contain lots of attributes which results in high workload during analysis. In order to improve the efficiency of dataset it needs to convert high dimensional dataset into low dimensional dataset. In data selection, some attributes are selected from existing dataset which are more relevant. In data extraction, a new feature set is extracted from existing dataset which is reduced dataset. Data analysis is to analyze the data which can be used to get useful patterns from raw data. Now a days, data mining is widely used in healthcare cooperation for exploring hidden patterns from dataset which is better used for disease prediction. Other programs of records mining encompass financial banking, manufacturing engineering, fraud detection, market basket analysis, CRM, intrusion detection, customer segmentation, corporate surveillance, crook investigation, lie detection, patron segmentation, corporate surveillance, crook investigation, bio-informatics and many more [1,2].

Rest of the paper is organized as follows, Section II contains the introduction of K-means clustering algorithm and how it is used as dimensionality reduction, Section IV contain the related work of disease prediction and data mining, section V explains the methodology of proposed work with flow chart, Section VI describes results and discussions of classification, and Section VII concludes conclusion of this work.

II. K-MEANS CLUSTERING

K-means clustering is an unsupervised machine learning algorithm which means clustering of dataset is done without

any class label. The dataset is clustered into k number of clusters in which k must be specified by the user. Data points within a cluster are similar to each other and dissimilar to the data points of other clusters. Each cluster must contain at least one data point. Clustering is done by choosing k number of random centroids, then finding Euclidean distance between centroids and all the data points in dataset. Centroids are also known as cluster centers. Value of k must be a positive integer and less than the number of data objects in dataset. A data point is assigned to that cluster in which centroid has smallest distance from the data point. After all the data points are assigned to one of the clusters the centroid is recalculated by finding mean of all the data points in particular cluster. Again, Euclidean distance is calculated between centroids and data points then the reassignment of data points is performed. The procedure is repeated till identical clusters are determined [3,4].

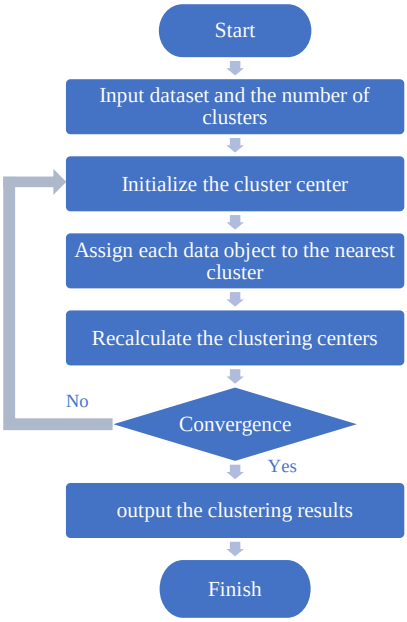

Figure 1. Flow chart of k means clustering

A. K-means Clustering Algorithm

Input: Dataset with d dimensions, k value.
Output: k clusters
Algorithm Steps:
Step1. Choose k initial centroids from the given dataset.
Step2. Find Euclidean distance between each centroid and all the data points in the dataset using following formula.

$$e(c_i,x_j)=\sqrt{(c_{i1}-x_{j1})^2+(c_{i2},x_{j2})^2+....+(c_{im},x_{jm})^2} \quad (1)$$

Step3. Assign each data point to the nearest cluster using above formula.
Step4. New centroids are calculated after assignment of every data point. Centroid recalculation is done on the basis of mean value of all data points within a cluster.
Step5. Step2 to 4 are repeated until no movement of data points are performed.

B. K-means clustering as Dimensionality Reduction

Due to hundreds of features in dataset the computation workload increases when this huge amount of data is processed. So, dimensionality reduction process can be used to decrease the computational workload on this data, by reducing the features. Dimensionality reduction is the process of reducing the features or attributes of dataset by extracting or selecting relevant features from dataset. Dimensionality reduction is done so that computational workload decreases and reduced features do not affect the classification results. Dimensional reduction includes two methods:

1. **Feature Selection:** Feature selection is the process of selecting the relevant features from the existing dataset.
2. **Feature Extraction**: Feature Selection is the process of extracting new features from the existing dataset by processing data.

K-means clustering algorithm can also be used as dimensionality reduction method. In this method feature extraction is performed by clustering similar features into one cluster and dissimilar features into other clusters. Input to the algorithm is dataset and number of clusters. Next, transform the dataset into transpose matrix. Clustering is performed on all the features or attributes of dataset. After applying k-means clustering on transpose matrix, the result of clustering is k number of features which is reduced feature set [5].

III. SUPPORT VECTOR MACHINE

Support vector machine is a machine learning algorithm that is used to classify the dataset into wide variety of class labels with the assist of hyper plane. Hyper plane is a line in vector space and a decision boundary that separates the data into two different types of classes. In SVM training and testing phase is there. In training phase, the model is trained using the class labels and feature values. Testing is done using this trained model. The hyper plane is selected on the basis of the margin width between two classes. The line that well separates two classes and has maximum vector margin is chosen as optimal hyper plane [6].

IV. RELATED WORK

Dona Sara Jacob *et al*. intends to provide a survey on breast cancer prediction with various data mining techniques. They compare various clustering and classification algorithms based on confusion matrix. Comparison is done between eight algorithms, which are C5.0, KNN, Naive Bayes, SVM, k-means, EM, PAM, Fuzzy c-means. They conclude that classification algorithms show better results than clustering algorithms [7].

M Mufli Muzakki *et al*. discovered an approach using k-means clustering algorithm and SVM algorithm to predict the dengue haemorrhagic fever in Bandung Regency. Dengue fever is of two types, normal dengue fever and dengue haemorrhagic fever. Dataset of DHF is collected from health department of Bandung Regency. Another dataset is weather dataset with six attributes which is taken from BMKG. K-means clustering algorithm is used as pre-processing to get high DHF and low DHF incident. To perform classification SVM is applied with dot kernel and radial to get accurate results [8].

Sarath Babu *et al.* has proposed an approach for heart disease prediction using genetic algorithm, k-means clustering algorithm, MAFIA algorithm and decision tree classification. Fourteen features from medical profile are used for heart disorder prediction. In this research work genetic algorithm is applied to extract the features from huge feature set. K-means clustering algorithm is performed on dataset with two features to get two clusters of high risk and low risk of heart disease. To figure out the most frequent patterns from correctly clustered dataset MAFIA algorithm is applied. Decision tree is applied in this paper for classification of heart disease which provides efficient results [9].

Mani Shankar *et al.* exposed a novel method for disease recognition based on patient's symptoms. They create their own dataset by collecting medical information from students of their college campus. Dataset includes students with name of the disease which they had suffered from along with all the symptoms of disease with severity rating and cure time of disease. Disease is recognized by calculating cut off value using formulae which includes all the symptoms and their severity score. They used reinforcement learning approach to predict the cure time of disease [10].

Naganna Chetty *et al.* described two approaches for disease prediction. In first approach data is clustered using fuzzy c means clustering algorithm and data classification is done using KNN. In second approach fuzzy c means and fuzzy KNN is used to perform the clustering and classification respectively. Both approaches are applied on PIMA diabetes dataset and liver disorder dataset. The results show that the fuzzy c means clustering along with fuzzy KNN classification has high accuracy than fuzzy c means clustering along with KNN classification [11].

Tejaswini U. Mane proposed a hybrid approach to predict heart disease using improved k-means clustering algorithm and ID3 algorithm on big data. In basis k-means clustering, centroids are randomly selected. They concluded that improved k-means clustering gives better accuracy in selecting the centroids of clusters than simple k-means clustering algorithm. They also learnt about various attributes which affects the heart [12].

V. METHODOLOGY

In our work breast cancer dataset is collected from UCI repository with nine attributes and a class label. The dataset is refined by removing the tuples containing missing values. After refinement of dataset, feature extraction is done by using k-means clustering algorithm. Few limitations of K-means clustering are, user have to assign value of k in advance and have to choose random initial centroids. In order to overcome the drawbacks of k-means clustering algorithm, an improved k-means clustering algorithm is discovered. To choose optimal value of k, the silhouette plot is used. Silhouette value is used in k-means clustering for identifying the number of clusters. Silhouette values are plotted on 2D plane.

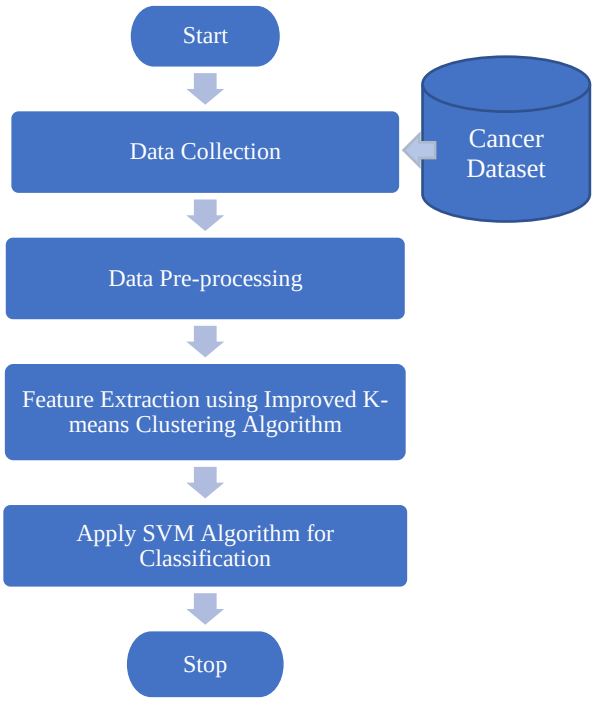

Figure 2. Flow chart of Proposed Work

Silhouette value of every data point for different number of clusters is plotted. For different number of clusters values, average silhouette value of all data points is computed. To find the value of k, average silhouette value is calculated and the cluster which results the highest silhouette value is assigned to k. After finding the value of k, the initial centroids are computed by partitioning the dataset into k equal parts and taking the average mean value as initial centers. Next, apply basic k-means clustering algorithms to extract the features from dataset. The reduced features are input to the SVM algorithm which classifies the dataset into

benign and malignant class. The implementation of proposed work is performed in MATLAB platform.

A. Data collection

Dataset in our research is Breast Cancer Wisconsin Original dataset which is taken from UCI repository. Dataset contains nine attributes, 699 instances with some missing values. Dataset attributes are Clump Thickness, Cell Size, Cell Shape, Marginal Adhesion, Normal Nucleoli, Single Epithelial Cell, Bare Nuclei, Bland Chromatin and Mitoses. All the above attributes have values between the range 1 to 10. Class label attribute has either benign class or malignant class which are represented as 2 and 4 respectively.

	clump thickness	uniformity of cell size	uniformity of cellshape	margina adhesion	single epithelial cell size	bare nuclei	bland chromatin	normal nucleoli	mitosis	class
1										
2	5	1	1	1	2	1	3	1	1	2
3	5	4	4	5	7	10	3	2	1	2
4	3	1	1	1	2	2	3	1	1	2
5	6	8	8	1	3	4	3	7	1	2
6	4	1	1	3	2	1	3	1	1	2
7	8	10	10	8	7	10	9	7	1	4
8	1	1	1	1	2	10	3	1	1	2
9	2	1	2	1	2	1	3	1	1	2
10	2	1	1	1	2	1	1	1	5	2
11	4	2	1	1	2	1	2	1	1	2
12	1	1	1	1	1	1	3	1	1	2
13	2	1	1	1	2	1	2	1	1	2
14	5	3	3	3	2	3	4	4	1	4
15	1	1	1	1	2	3	3	1	1	2
16	8	7	5	10	7	9	5	5	4	4
17	7	4	6	4	6	1	4	3	1	4
18	4	1	1	1	2	1	2	1	1	2
19	4	1	1	1	2	1	3	1	1	2
20	10	7	7	6	4	10	4	1	2	4
21	6	1	1	1	2	1	3	1	1	2
22	7	3	2	10	5	10	5	4	4	4
23	10	5	5	3	6	7	7	10	1	4
24	3	1	1	1	2	1	2	1	1	2
25	8	4	5	1	2	?	7	3	1	4
26	1	1	1	1	2	1	3	1	1	2

Figure 3. Dataset with missing values

B. Data Pre-processing

Pre-processing is to refine the data by data cleaning and data reduction. Data cleaning is to clean the dataset so there is no missing value in the dataset. Missing values can be removed by either deleting the whole row which has missing value or filling the missing value with the average mean value of column. In case of a smaller number of instances in dataset the missing value rows are not deleted rather it fills with mean values. Data reduction is the process of reducing the data or reducing the dimensionality of the dataset by using the k-means clustering algorithm. Dimensionality reduction is essential for the reduction of computational workload on classifier and to achieve better results.

	clump thickness	uniformity of cell size	uniformity of cellshape	margina adhesion	single epithelial cell size	bare nuclei	bland chromatin	normal nucleoli	mitosis	class
1										
2	5	1	1	1	2	1	3	1	1	2
3	5	4	4	5	7	10	3	2	1	2
4	3	1	1	1	2	2	3	1	1	2
5	6	8	8	1	3	4	3	7	1	2
6	4	1	1	3	2	1	3	1	1	2
7	8	10	10	8	7	10	9	7	1	4
8	1	1	1	1	2	10	3	1	1	2
9	2	1	2	1	2	1	3	1	1	2
10	2	1	1	1	2	1	1	1	5	2
11	4	2	1	1	2	1	2	1	1	2
12	1	1	1	1	1	1	3	1	1	2
13	2	1	1	1	2	1	2	1	1	2
14	5	3	3	3	2	3	4	4	1	4
15	1	1	1	1	2	3	3	1	1	2
16	8	7	5	10	7	9	5	5	4	4
17	7	4	6	4	6	1	4	3	1	4
18	4	1	1	1	2	1	2	1	1	2
19	4	1	1	1	2	1	3	1	1	2
20	10	7	7	6	4	10	4	1	2	4
21	6	1	1	1	2	1	3	1	1	2
22	7	3	2	10	5	10	5	4	4	4
23	10	5	5	3	6	7	7	10	1	4
24	3	1	1	1	2	1	2	1	1	2
25	1	1	1	1	2	1	3	1	1	2
26	5	2	3	4	2	7	3	6	1	4

Figure 4. Dataset with no missing values

C. Data Classification

After most relevant features are extracted from the dataset, those features are provided as input to the SVM classifier which builds a model using 80% of whole dataset as training data. Then test data which is remaining 20% of pre-processed dataset given as input to the model classifies the dataset into two classes called benign and malignant type of tumour.

VI. Results and Discussion

In machine learning, the error rate of a training model should be low and accuracy should be high, in order to get a smaller number of misclassifications. To get better results with high accuracy improved k-means clustering is defined in this paper which extracts the correct features from given feature set. According to silhouette plot results, value of k is defined. Silhouette plot is basically used for k-means clustering for finding the correct value of k by calculating the silhouette score for each instance. The silhouette plot which gives highest average silhouette score for a particular cluster size is selected as k value. Figure 5 shows the silhouette plot for 3 clusters. For 3 clusters the average silhouette value is highest, so the value of k is 3. After value of k is defined, the features of dataset are divided into 3 clusters using k-means clustering. From these three features two optimal features are given as input to SVM for training purpose. In figure 6 and 7, extracted features are represented along x & y axis and benign tumour is represented as '0' and malignant tumour is represented as '+' sign.

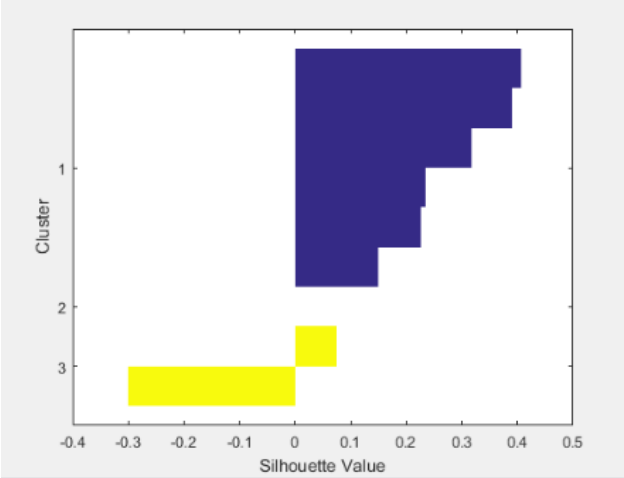

Figure 5. Silhouette Plot for 3 Clusters

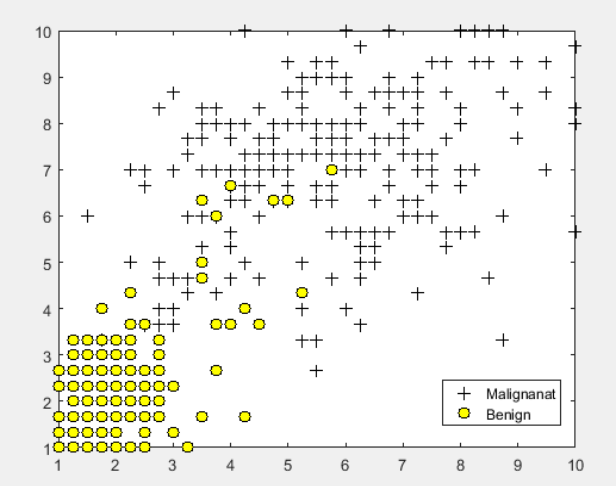

Figure 6. Benign and Malignant tumour

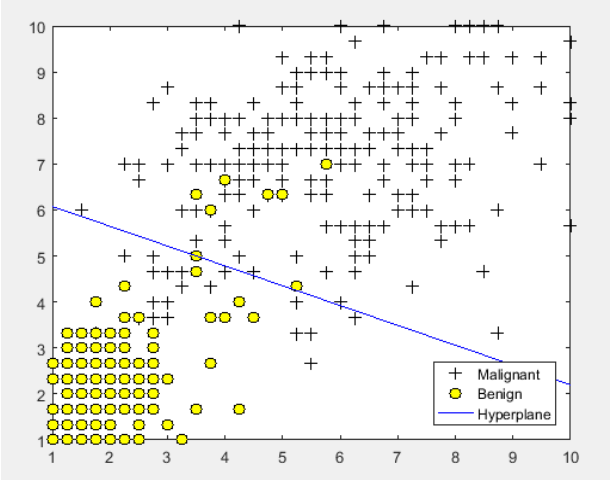

Figure 7. Classification of cancer dataset using SVM

Figure 7 shows the classification of cancer dataset using SVM in which separation is represented by hyperplane. After the model is trained with SVM, a line called hyperplane is selected which has largest margin and drawn between two classes. The actual class labels are represented by '0' and '+' symbols but using SVM model the decision boundary predicts that the instances below the hyperplane are benign and above the hyperplane are malignant which are called as predicted classes. Prediction is performed under the testing phase with the test data. Accuracy of classifier is calculated by following formula:

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

Where, TP denotes true positive results, TN denotes true negative results, FP denotes false positive results and FN denotes false negative results. True positive results are the number of positive instances which are correctly predicted by the classifier. True negative results are the number of negative instances that are correctly predicted by the classifier. False positive results are the number of negative instances which are incorrectly predicted as positive instances by the classifier. False negative results are the number of positive instances which are incorrectly predicted as negative instances by the classifier.

VII. CONCLUSION

In this paper, improved k-means clustering algorithm and SVM classification algorithm has been proposed for breast cancer disease prediction. The proposed work incorporates improved k-means clustering algorithm for dimensionality reduction to extract attributes. On the contrary, simple k-means clustering algorithm takes user defined k value in clustering and selection of initial clusters are random. In improved k-means clustering algorithm, value of k is defined by using silhouette plot. The results of SVM algorithm using improved k-means clustering gives more accuracy as compared to simple k-means.

Acknowledgment

The author is highly thankful to the faculty of Guru Nanak Dev Engineering College, Ludhiana.

References

[1] M. Umamaheswari, P.I. Devi, "*Prediction of myocardial infarction using K-medoid clustering algorithm*," In the Proceedings of the 2017 IEEE International Conference Intelligent Techniques in Control, Optimization and Signal Processing, Srivilliputhur, **India**, pp.**1-6, 2018.**

[2] Nisha, P.J. Kaur, "*A Survey of Clustering Techniques and Algorithms*," In the Proceedings of the 2015 International Conference on Computing for Sustainable Global Development, New Delhi, **India**, pp.**304–307, 2015.**

[3] P. Manivannan, P.I. Devi, "*Dengue fever prediction using K-means clustering algorithm*," In the Proceedings of the 2017 IEEE International Conference Intelligent Techniques in Control, Optimization and Signal Processing, Srivilliputhur, **India**, pp.**1-5, 2018.**

[4] V. Jain, "*Outlier Detection Based on Clustering Over Sensed Data Using Hadoop*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**1**, Issue.**2**, pp.**45-50**, **2013**.

[5] P.S. Bishnu, V. Bhattacherjee, "*A dimension reduction technique for K-Means clustering algorithm*," In the Proceedings of the 2012 1st International Conference on Recent Advances in Information Technology, Dhanbad, **India**, pp.**531–535, 2012.**

[6] J. Meena, A. Mandloi, "*Classification of Data Mining Techniques for Weather Prediction*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**4**, Issue.**1**, pp.**21-24**, **2016**.

[7] D.S. Jacob, R. Viswan, V. Manju, L. Padmasuresh, S. Raj, "*A Survey on Breast Cancer Prediction Using Data Mining Techniques*," In the Proceedings of the 2018 Conference on Emerging Devices and Smart Systems, Tiruchengode, **India**, pp.**256-258, 2018.**

[8] M.M. Muzakki and F. Nhita, "*The spreading prediction of Dengue Hemorrhagic Fever (DHF) in Bandung regency using K-means clustering and support vector machine algorithm*," In the Proceedings of the 2018 6th International Conference on Information and Communication Technology, Bandung, **Indonesia**, pp.**453–458, 2018.**

[9] S. Babu *et al.*, "*Heart disease diagnosis using data mining technique*," In the Proceedings of the 2017 International Conference of Electronics, Communication and Aerospace Technology, Coimbatore, **India**, pp.**750–753, 2017.**

[10] M. Shankar, M. Pahadia, D. Srivastava, T.S. Ashwin, and G.R.M. Reddy, "*A Novel Method for Disease Recognition and Cure Time Prediction Based on Symptoms*," In the Proceedings of the 2015 2nd Internationl Conference on Adavnces in Computing and Commuication Engineering, Dehradun, **India**, pp.**679–682, 2015.**

[11] N. Chetty, K. S. Vaisla, and N. Patil, "*An Improved Method for Disease Prediction Using Fuzzy Approach*," Proc. - 2015 2nd International Conference on Advances in. Computing Communication Engineering, Dehradun, **India**, pp.**568–572, 2015.**

[12] T.U. Mane, "*Smart heart disease prediction system using Improved K-means and ID3 on big data*," 2017 International Conference on Data Management, Analytics and Innovation, Pune, **India**, pp.**239–245, 2017.**

Authors Profile

C Kaur studied Bachelor of Technology in Computer science and engineering from Guru Nanak Dev Engineering college, Ludhiana (Punjab) in 2017. She is currently pursuing Master of Technology in Computer Science & engineering from Guru Nanak Dev Engineering College, Ludhiana (Punjab). Her main research focuses on development of disease prediction system with improved k-means clustering algorithm and SVM algorithm to improve performance.

K Sharma received a Bachelor degree in Computer Science Engineering in 2011 and a Master degree in Computer Science & Engineering in 2013. He is currently pursuing phd in Computer Science & Engineering. He is with Guru Nanak Dev Engineering College, Ludhiana (Punjab) India as Assistant Professor, in CSE department since 2015- 2019. His research area is data mining and big data.

Amandeep K. Sohal received a Bachelor degree in Electrical Engineering from Punjab Technical University, Jalandhar (Punjab) India in 2001 and a Master degree in Computer Science & Engineering from Punjabi University Patiala, India, in 2004. She is with Guru Nanak Dev Engineering College, Ludhiana (Punjab) India as Assistant Professor, in CSE department since 2004- 2019. Her research area is data communications, computer networks, wireless communications, wireless sensor networks and machine learning.

