International Journal of Advanced Technology and Engineering Exploration, Vol 6(53)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2019.650034

Research Article

An efficient SKNN based approach for heart disease classification

Heena Farheen Ansari[1*] and Varsha Namdeo[2]
Research Scholar, Computer Science and Engineering, Sarvepalli Radhakrishnan University, Bhopal, MP, India[1]
Associate Professor, Computer Science and Engineering, RKDFIST Bhopal, MP, India[2]

Received: 15-February-2019; Revised: 20-April-2019; Accepted: 25-April-2019

Abstract
An efficient span-k-nearest neighbour (SKNN) algorithm has been proposed. It is used for the categorization of heart disease. The objective is to differentiate the data and find the accuracy of detection. The pre-processing is done based on the three attributes combination that is two, three and four with the help of KNN method. Then it is categorized based on five different spans that are 100, 125, 150, 200 and 250. The proposed work is compared with different factors of SKNN so that the proper capability can be explored. The obtained results show that the proposed approach has the significant capability of better classification.

Keywords
Heart disease, Data mining, KNN, Data classification.

1.Introduction
There are higher death rates recorded in case of heart diseases worldwide [1]. The other report also indicates the warning condition in India [2]. This condition is recorded from several repositories available worldwide. Taking these reports as the base the direction of the paper in the detection of the heart disease.

This paper provides the discussion in comparative way which includes the methodology discussion, including clustering and classification techniques [2−9]. Different aspects to find out the betterment is analysed and discussed for the approach to analyse the prospective of the data to handle and analysing the data with the attributes to finding better data categorization [10−15].

To clear up and analyse the use data mining to highlight to discover discovering that isn't simply correct, yet moreover fathomable for the diverse viewpoints [16]. Understandability is imperative at whatever point discovered learning will be used for supporting a human decision [17].

*Author for correspondence

Everything considered, whenever discovered data isn't understandable for the information and the extraction, it won't be possible to disentangle and support the learning [18].

Data mining is the nontrivial extraction of undeniable, officially dark, and possibly accommodating information from data. This incorporates different unmistakable specific strategies, for instance, batching, data layout, learning request rules, finding dependence frameworks, dismembering changes, and perceiving variations from the norm [19]. The found data and learning are helpful for different applications, including market examination, choice, strengthen, misdirection territory, and business association. Different methodologies have been proposed to think data, and mining restorative information is an essential field [20].

2.Literature review
In 2017, Burse and Wadhvani [20] suggested that the mortality rate is highly increased by heart disease. They have suggested different heart disease contributing factors like heart-rate, cholesterol levels, blood pressure, etc. They have reviewed Logistic Regularization, Lasso, Elastic Net and Group Lasso regularization techniques. Different subsets of

features have been selected based on the regularization.

In 2017, Devi et al. [21] suggested that the sudden cardiac death is the major issue of death because of the cardiac reasons. The authors have tried to predict the sudden cardiac death one hour earlier. For the experimentation ECG signals have been considered from the online database. They have considered the QRS peaks in the heart rates for the consideration and investigation.

In 2017, Geetha et al. [22] suggested that the online transaction rates now days have been increased. They have suggested and analyze different data mining techniques for the analysis of the knowledge discovery from the databases. These techniques have also been analyzed in the health sector.

In 2017, Gnaneswar and Jebarani [23] suggested that there is enormous development has been noted in the health care center. Because of the heart disease, there are several deaths has been noted in the world-wide scenario. They have provided the methodological analysis and survey based on the methodology pros and cons.

In 2017, Islam et al. [24] suggested that the hazardous position of heart diseases worldwide. So, they have suggested earlier detection can help. They have suggested the photoplethysmography (PPG) signal for the heart disease detection. They have suggested that there may be tissue damage due to the traditional diagnosis. It is extracted from human facial video. According to the authors it is also efficient and cost effective.

In 2017, Mahboob et al. [25] suggested that the heart disease detection is important. They have considered hidden Markov models, support vector machine, feature selection, computational intelligent classifier, prediction system, data mining techniques and genetic algorithm.

In 2017, Pahwa and Kumar [26] suggested that the hidden pattern discovery is an important task. They have suggested the need of advanced data mining technique in case of medical sector. They have used SVM-RFE method and gain ratio algorithm. Their results show the efficiency of the approach.

3.Methods

In this paper, we have applied SKNN method which is based on KNN. It is efficient in storing all the possible cases and classifies new cases based on a similarity measure. The first data is loaded, and then pre-processing is performed based on the attribute and the span. The span considered here are 100, 125, 150, 200 and 250. It provides the data arrangement based on the different nearest attribute weights. Then it is processed based on span and iterated till the following span and generated the grouping. The span provides us the more iterative filtration for pattern classification and categorizing, as it provides a deep filtering with different class labels. This provides a type of classification which is capable to determine the alike span groups. The group's size proportion imperative forces a requirement on the yield such that group size for all categories in the subsequent arrangement of groups. Substantial estimations where it is nearer to the disease will assign 0 otherwise the estimation will provide 1. It supports the calculation of grouping in the classification of alike class. The size proportion imperative is upheld in an extremely gullible style, by irregular reseeding. Since this can be a somewhat tedious procedure, it is conceivable to set an upper bound on the quantity of reseeds done by the calculation. This upper bound is characterized by the parameter maximum number of reseeds. It ought to be noted however that typically there's no compelling reason to utilize the size, proportion imperative, selecting the coveted number of bunches will, by and large, bring about groups of generally equivalent size, given all around appropriated information. Then final accuracy has been calculated based on the achieved weights.

The span-KNN algorithm is shown below:
Notations:
The values considered are numerical value. The attributes considered are 2, 3, 4, 5 etc.
W_i: Random weight
X_i: Instance
N: Span

Algorithm 1: SKNN
Step 1: Accept the data from the whole set.
Step 2: The value of k is set according to the attribute.
Step 3: Set the learning rate and span of iterations.
Step 4: Assign W_i to each instance X_i.
Step 5: Iterations following N span. In our case it is 5. These are 100, 125, 150, 200 and 250.
Step 6: Each set in K for example N_k in N, do n Set N_k = validation set n
Step 7: For every X_i in N such that X in N, X_i does not belong to N_k is not considered as the neighbour.

Step 8: Then K nearest neighbours are calculated based on Euclidean distance.
Step 9: Calculate the class value
$\sum W_k X X_{j,K}$ where j is the class attribute n
Step 10: If actual class! = predicted class then the value is discarded
Step 10: Calculate the accuracy as = (# of correctly classified sets / # set in N_k) X 100

4.Results

In our approach we have used Statlog heart disease dataset. This dataset is a collection of total 13 attributes and one decision attributes. There are total 270 records. The variables to be classified are 1 and 2. Then Cleveland heart disease dataset have been used. This dataset is a collection of 13 attributes and one decision attributes. There are total 303 records. The variables to be classified are 1 and 2. In this section the results have been discussed with the comparative study presented in the comparison. The proposed work is compared with different factors like attributes and five different epochs like 100, 125, 150, 200and 250. *Figure 1-3* shows the result with a KNN method with random attribute value 2, 3, 4 respectively in case of Statlog dataset. *Figure 4-6* shows the result with a KNN method with random attribute value 2, 3, 4 respectively in the case of Cleveland dataset.

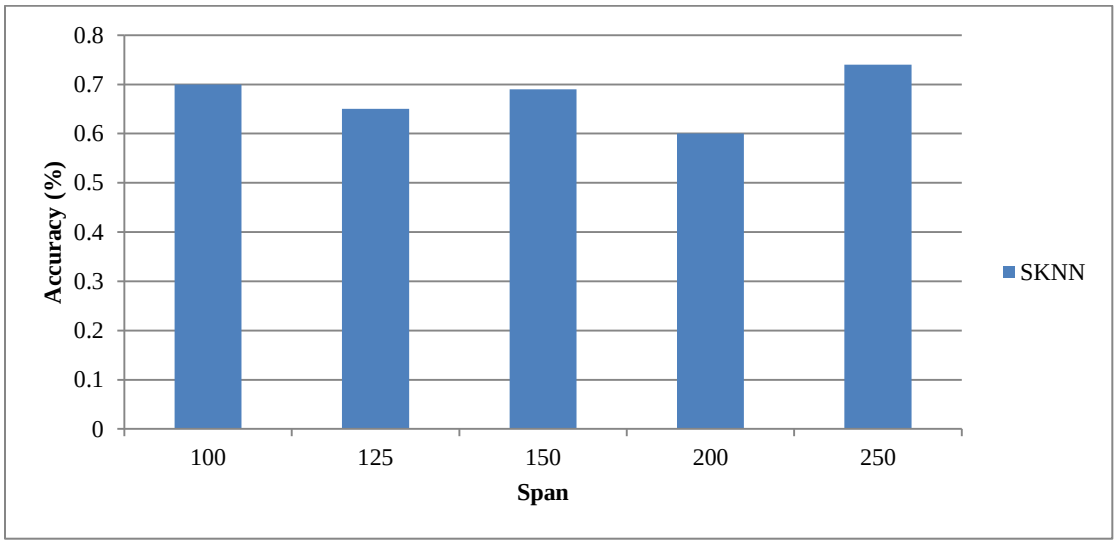

Figure 1 Result with KNN method with random attribute 2 (Statlog)

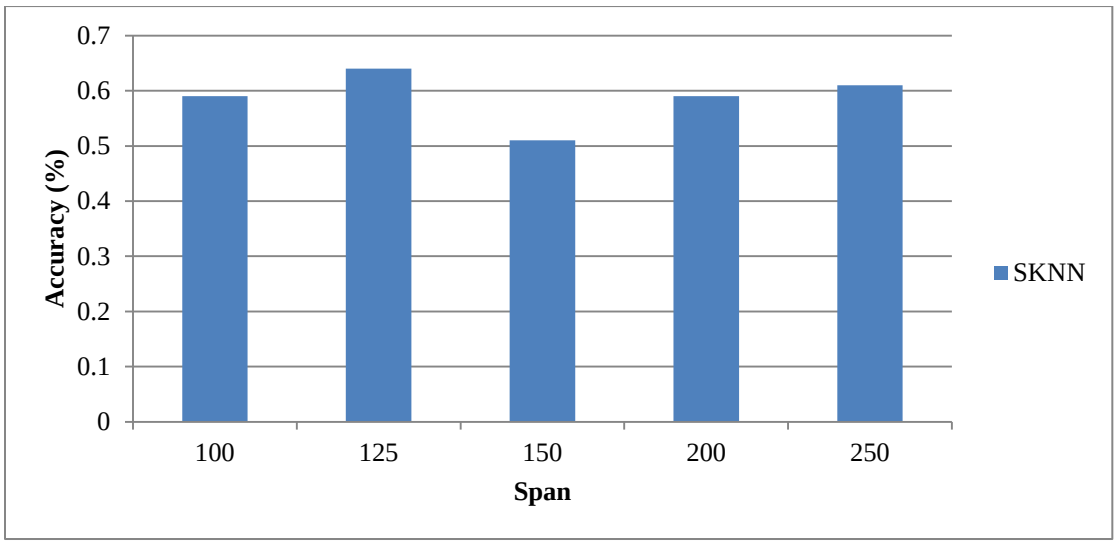

Figure 2 Result with KNN method with random attribute 3 (Statlog)

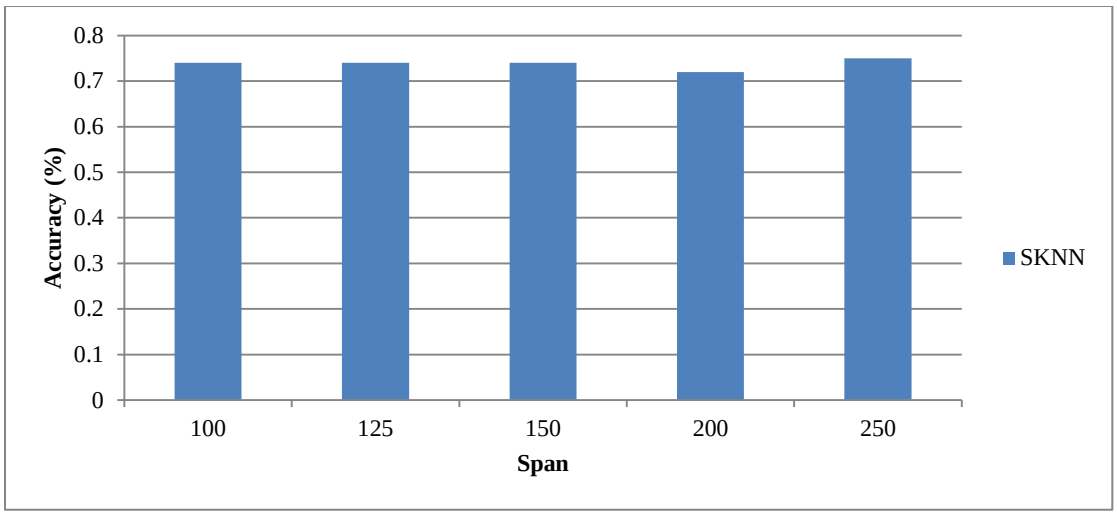

Figure 3 Result with KNN method with random attribute 4 (Statlog)

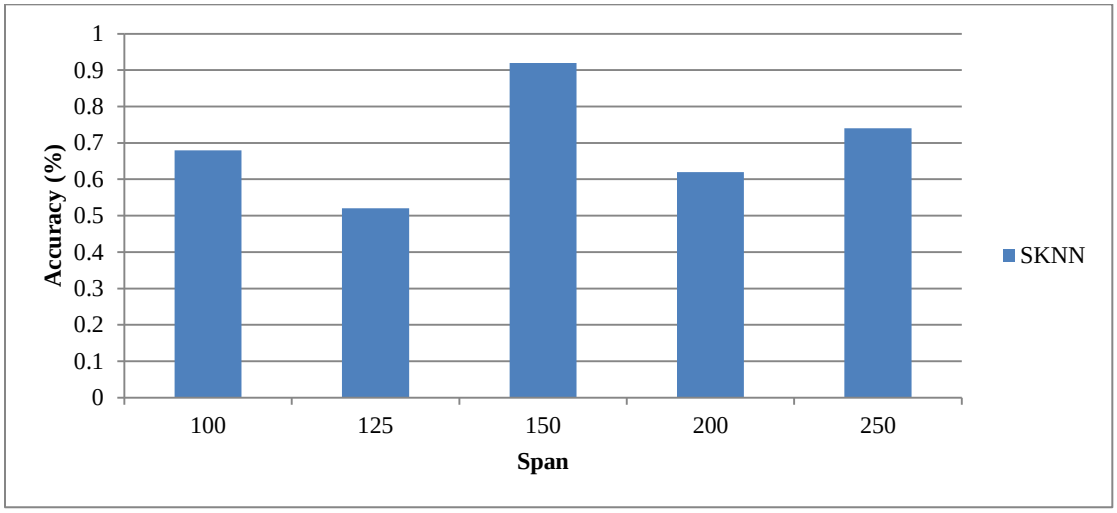

Figure 4 Result with KNN method with random attribute 2 (Cleveland)

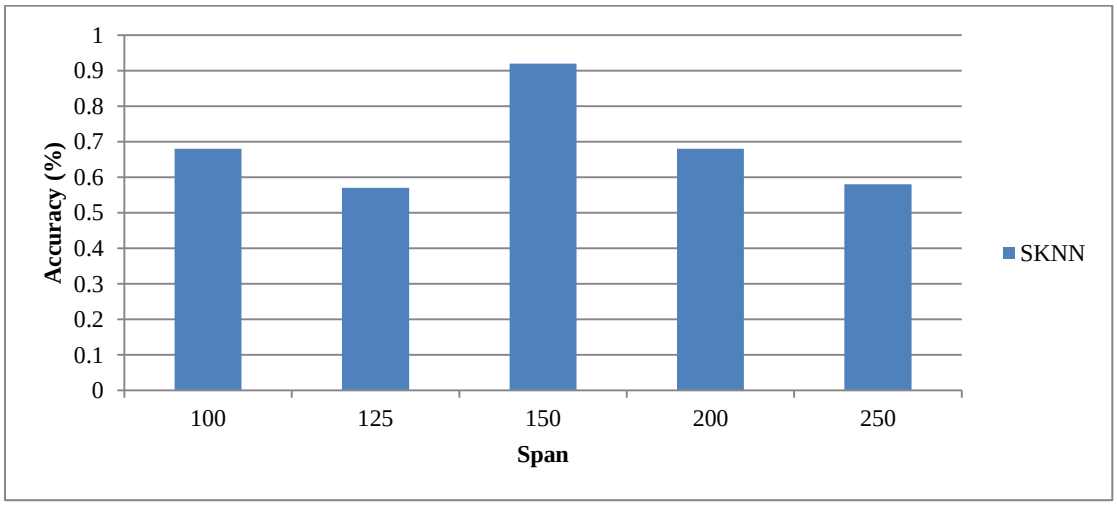

Figure 5 Result with KNN method with random attribute 3 (Cleveland)

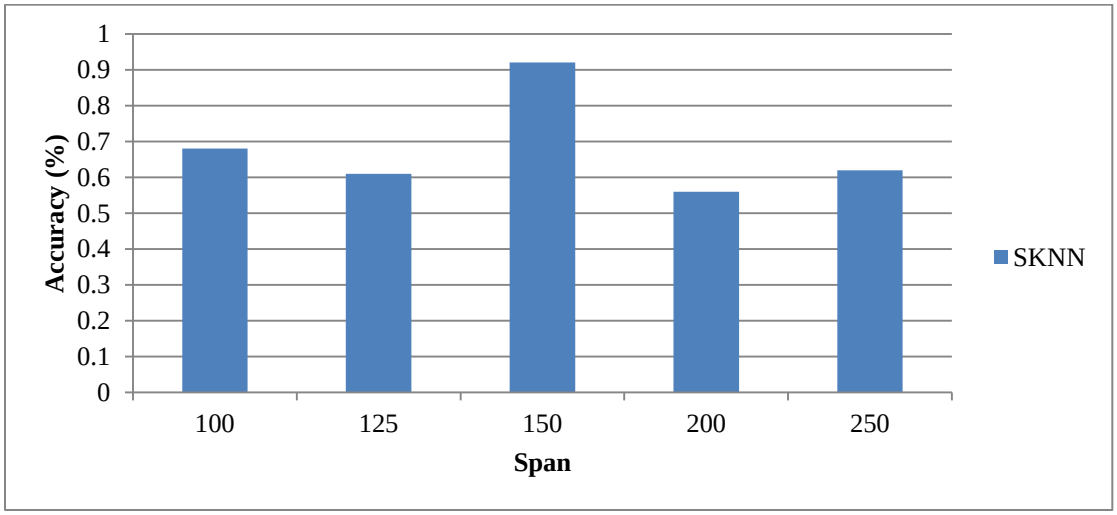

Figure 6 Result with KNN method with random attribute 4 (Cleveland)

5.Conclusion

In this paper an efficient mechanism based on span-KNN (SKNN) has been applied for better heart disease classification. The dataset is considered from the UCI repository. Two datasets are considered for the experimentation and comparison. These are Statlog heart disease dataset and Cleveland heart disease dataset. Data is categorized with KNN with five different spans with 100, 125, 150, 200 and 250. This category is used for deep filtration and classification. It is shown from the results that the SKNN has the capability to perform better classification.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Dubey AK, Choudhary K. A systematic review and analysis of the heart disease prediction methodology. International Journal of Advanced Computer Research. 2018; 8(38):240-56.

[2] Prabhakaran D, Jeemon P, Roy A. Cardiovascular diseases in India: current epidemiology and future directions. Circulation. 2016; 133(16):1605-20.

[3] Dubey AK, Gupta U, Jain S. Breast cancer statistics and prediction methodology: a systematic review and analysis. Asian Pacific Journal of Cancer Prevention. 2015; 16(10):4237-45.

[4] Dubey AK, Gupta U, Jain S. A survey on breast cancer scenario and prediction strategy. In proceedings of the international conference on frontiers of intelligent computing: theory and applications 2015 (pp. 367-75). Springer, Cham.

[5] Dubey AK, Gupta U, Jain S. Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis. Chinese Journal of Cancer. 2016; 35(1).

[6] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining.

[7] Kushwah J, Singh D. Classification of cancer gene selection using random forest and neural network based ensemble classifier. International Journal of Advanced Computer Research. 2013; 3(10):30-4.

[8] Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International Journal of Computer Assisted Radiology and Surgery. 2016; 11(11):2033-47.

[9] Peter TJ, Somasundaram K. An empirical study on prediction of heart disease using classification data mining techniques. In international conference on advances in engineering, science and management 2012 (pp. 514-8). IEEE.

[10] Li K, Li P. A selective fuzzy clustering ensemble algorithm. International Journal of Advanced Computer Research. 2013; 3(13):1-6.

[11] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. In CSI sixth international conference on software engineering 2012 (pp. 1-6). IEEE.

[12] Kumar S, Patidar K, Kushwah R, Chouhan S. A review and analysis on text data encryption techniques. International Journal of Advanced Technology and Engineering Exploration. 2017; 4(30):88-92.

[13] Gawade P, Chauhan RP. Detection of lung cancer using image processing techniques. International Journal of Advanced Technology and Engineering Exploration. 2016; 3(25):217-22.

[14] Singh S, Yadav M, Gupta H. Finding the chances and prediction of cancer through Apriori algorithm with transaction reduction. International Journal of Advanced Computer Research. 2012; 2(4):23-8.
[15] Elkader SA, Elmogy M, El-Sappagh S, Zaied AN. A framework for chronic kidney disease diagnosis based on case based reasoning. International Journal of Advanced Computer Research. 2018; 8(35):59-71.
[16] Li K, Gao Y. Fuzzy clustering with the generalized entropy of feature weights. International Journal of Advanced Computer Research. 2016; 6(27):195-208.
[17] Shengbing C, Xiaofeng W, Xiaofang W. Mining dynamical frequent itemsets based on ant colony algorithm. In international conference on computer science and automation engineering 2011 (pp. 252-5). IEEE.
[18] Vaska JS, Sowjanya AM. Clustering diabetics data Using M-CFICA. International Journal of Advanced Computer Research. 2015; 5(20):327-33.
[19] Jamil A, Salam A, Amin F. Performance evaluation of top-k sequential mining methods on synthetic and real datasets. International Journal of Advanced Computer Research. 2017; 7(32):176-84.
[20] Burse R, Wadhvani R. Comparative analysis of various regularization techniques in the prediction of heart diseases. In international conference on inventive computing and informatics 2017 (pp. 356-61). IEEE.
[21] Devi R, Tyagi HK, Kumar D. Early stage prediction of sudden cardiac death. In international conference on wireless communications, signal processing and networking 2017 (pp. 2005-8). IEEE.
[22] Geetha MC, Shanthi IE, Sehnaz NS. Analyzing the suitability of relevant classification techniques on medical data set for better prediction. In international conference on I-SMAC (IoT in social, mobile, analytics and cloud) 2017 (pp. 665-70). IEEE.
[23] Gnaneswar B, Jebarani ME. A review on prediction and diagnosis of heart failure. In international conference on innovations in information, embedded and communication systems 2017 (pp. 1-3). IEEE.
[24] Islam M, Ashikuzzaman M, Tabassum T, Yusuf MS. A non-invasive technique of early heart diseases prediction from photoplethysmography signal. In international conference on electrical information and communication technology 2017 (pp. 1-5). IEEE.
[25] Mahboob T, Irfan R, Ghaffar B. Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics. In internet technologies and applications 2017 (pp. 110-5). IEEE.
[26] Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In international conference on electrical, computer and electronics 2017 (pp. 500-4). IEEE.

Heena Farheen Ansari is a PhD Scholar in Computer Science & Engineering. She holds a Masters in Technology in Computer Science & Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya University, Bhopal, Madhya Pradesh, India. Her research interests are Data Mining and Evolutionary Algorithms.
Email: heena_pearl3@yahoo.co.in

Varsha Namdeo is an Associate Professor in the Department of Computer Science and Engineering at SRK University, Bhopal, India. She is a Teacher and Researcher in the field of Computer Science and Information Technology. She earned her Master in Computer Application from Barkatullah University, Bhopal (M.P.) in 2000 and in Computer Science and Engineering from Barkatullah University, Bhopal (M.P.) in 2009 and PhD degree from Maulana Azad National Institute of Technology; Bhopal (M.P.) in 2015. She had a long career in teaching and research.