International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Decision Tree based Classification and Dimensionality Reduction of Cervical Cancer

Diksha, Dinesh Gupta

Abstract*: The data revolution in medicines and biology have increased our fundamental understandings of biological processes and determining the factors causing any disease, but it has also posed a challenge towards their analysis. After breast cancer, most of the deaths among women are due to cervical cancer. According to IARC, alone in 2012 a noticeable number of cases estimated 7095 of cervical cancer were reported. 16.5% of the deaths were due to the cervical cancer with the total deaths of 28,711 among women. To analyze the high dimensional data with high accuracy and in less amount of time, their dimensionality needs to be reduced to remove irrelevant features. The classification is performed using the recent iteration in Quinlan's C4.5 decision tree algorithm i.e. C5.0 algorithm and PCA as Dimensionality Reduction technique. Our proposed methodology has shown a significant improvement in the account of time taken by both algorithms. This shows that C5.0 algorithm is superior to C4.5 algorithm.

***Keywords*: Classification, Cervical Cancer, Decision Tree, Dimensionality Reduction.**

I. INTRODUCTION

Cervical cancer is the preeminent causes of fatality among women of age 30 or above. Cervical cancer is a tumor of the cervix which is the lowermost part of uterus [1]. Human Papillomavirus (HPV) is the virus which is responsible for the cervical cancer [1]. When the body's cervical cells growth is malignant, the extra cells form a tumor. Mostly women have good immunity against HPV infection but if not, it can lead to cancer. Cervical cancer does not show any symptoms in the beginning. Regular check-ups are needed to diagnose it. But if the cancer has been diagnosed at early stages, it can be easily cured through various treatments. Size of the tumor is the key factor in deciding the type of treatment that is best fit for individual cases. Cervical cancer when left undiagnosed can lead to vaginal bleeding, pelvic pain, and unusual vaginal discharge. Various risk factors lead to cervical cancer. Some of them are smoking cigarettes, taking contraceptive pills for many years, weak immune system, having HIV.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Diksha*, Department of Computer Science and Engineering, IKG Punjab Technical University, Jalandhar, India. Email: dikshaaggarwal100@gmail.com
Dinesh Gupta, Department of Computer Science and Engineering, IKG Punjab Technical University, Jalandhar, India. Email: dineshgupta@ptu.ac.in

The chance of cervical cancer becomes higher with old age [2]. Cervical cancer can also unfold to more body parts in three ways through tissue, the liquid body substance system, and the blood. Several cancer deaths are caused once cancer moves from the first growth and spreads to alternative tissues and organs. This can be referred to as pathologic process cancer. The 2010 WHO/ICO outline report states that girls aged 15 or above are prone to cervical cancer in Malaysia which has a population of 8.7 million of such girls. Annually, 631 out of 2126 girls suffering from cervical cancer die from the disease [3]. The American Cancer Society's estimates for cervical cancer in the U.S. for 2019 are [2] (1) Around 13,170 novel cases of cervical cancer will be analyzed, and (2) Around 4,250 women will expire from cervical cancer.
Many approaches have been proposed to detect the cervical cancer at early stages which are support vector machine (SVM), k-nearest neighbor, linear regression, and naive-bayes [1, 4]. Various dimensionality reduction techniques have been implemented in order to achieve low dimensional data space. The most common among the various used techniques is Principal Component Analysis (PCA). This algorithm is a useful statistical method which has its applications in various fields like speech recognition, image recognition [5, 6], text processing, and scientific data processing [7], and recommendation engines. Rest of the paper is formulated as follows, Section II contains the introduction to dimensionality reduction, Section III contains some introduction to the classification, Section IV contains the related work, section V describes the proposed procedure with flow chart, Section VI interpret results, and Section VIII concludes research work with future directions

II. DIMENSIONALITY REDUCTION

The growth in data increases the sparsity in the data and unnecessarily increases storage space and processing time to analyze and classify the data. This is termed as Curse of Dimensionality. Value brought by way of extra size is plenty smaller in comparison to overhead it adds to the algorithm. To reduce the overhead of higher dimensionality, researchers often extract specific attributes (i.e. features) that are relevant in drawing the results by scaling down the dimensionality of the data.

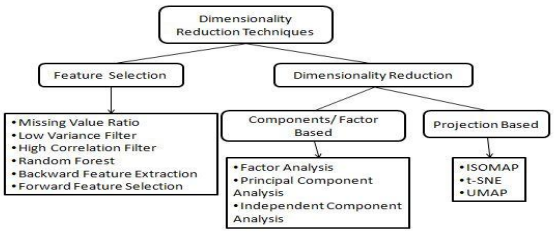

Figure 1: Dimensionality Reduction Techniques

III. CLASSIFICATION

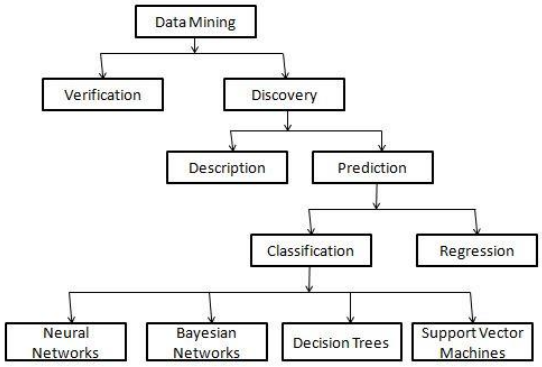

Figure 2: Data Mining Paradigms

Classification is an effective data mining tool to discover the hidden potential information in the database. This is also known as Knowledge Discovery in Databases (KDD). It is a supervised machine learning technique in which the unseen data samples are grouped into one of the known target classes. The different methodologies like decision tree, neural networks, support vector machine can be used for classifying the information.

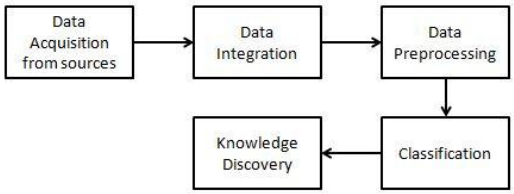

Figure 3: Classification Process.

IV. RELATED WORK

Maaten et al. [8] performed various dimensionality reduction techniques on different artificial and natural datasets. The techniques which do not employ graphs like PCA, auto-encoder performed well on the natural datasets. The non-linear dimensionality reduction techniques do not outperformed PCA.

Nasution et al. [9] implemented PCA as dimensionality reduction method to boost the performance of the decision tree C4.5 classification. Noisy data and irrelevant features causes overfitting and it has great impact on the splitting by the decision tree. So PCA algorithm is implemented first to reduce the noise in the data and to select the non-correlated features from the dataset of the cervical cancer. And then decision tree C4.5 algorithm was used for the classification. The dataset has 858 instances and 36 attributes with four target classes. The target classes were reduced into two target values signifying whether the patient has cervical cancer or not. Implementing PCA on the dataset has reduced the number of features from 36 to 12 with the increased accuracy of 90.7 percent.

Revathy et al. [10] has compared the performance statistics of the decision tree algorithms C4.5 and C5.0 on the dataset of crop pest data. Feature selection technique OneR has been used to extract the relevant features. While considering required parameters, this paper concludes that C5.0 comes out to be superior to other algorithm in machine learning. C5.0 supersedes C4.5 both in the terms of accuracy and computation time.

Patil et al. [11] compared the performance of CART and C5.0 algorithm to make decision for customers for recommending membership cards. CART selects its attribute upon which splitting will occur through GINNI Index whereas C5.0 uses Information Gain. The splitting criteria by both the algorithms differ slightly but C5.0 is efficient in accuracy than CART algorithm.

Agarwal et al. [12] also compared the two dimensionality reduction techniques PCA and LDA for the face recognition system. The paper depicted that PCA has better performance, speed and recognition technique than LDA.

Douangnoulack et al. [13] used reduced the dimensionality of the Wisconsin Breast Cancer using Principal Component Analysis and classified it with the C4.5 (J48) classification algorithm. PCA provides new features which greatly vary across the classes and model generated from these new features produced better results in classifying the data than the original features.

V. METHODOLOGY

A. Data Acquisition

The dataset of the cervical cancer from UCI repository is acquired from kaggle.com. The dataset has 858 instances and 36 attributes including 4 class attributes.

a. **About the dataset**

The dataset was compiled at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset includes demographic information, habits, and historic medical records of 858 patients. Due to some confidential concerns some patients did not responded to few questions which lead to missing values. The dataset is obtained from UCI Repository. This dataset contains risk factors for cervical cancer [14].

Table1: Cervical Cancer Dataset Attribute Description [14]

Sr. No.	Attribute Name	Sr. No.	Attribute Name
1	Age	**17**	Vulvo-perineal condylomatosis
2	Number of Sexual partners	**18**	Syphilis
3	Age of first sexual intercourse	**19**	Pelvic inflammatory disease
4	Num of pregnancies	**20**	Genital herpes
5	Smokes	**21**	Molluscum contagiosum
6	Smokes	**22**	AIDS
7	Smokes -packs/year	**23**	HIV
8	Hormonal Contraceptive	**24**	Hepatitis B
9	Hormonal Contraceptives –years	**25**	HPV
10	IUD	**26**	Number of diagnosis

11	IUD –years	**27**	Time since first diagnosis
12	STDs	**28**	Time since last diagnosis
13	STDs-number	**29**	Cancer
14	Condylomatosis	**30**	CIN
15	Cervical condylomatosis	**31**	HPV
16	Vaginal condylomatosis	**32**	Dx

Table2: Cervical Cancer Dataset Target Variables Description [14]

Sr. No.	Target Variable	Sr. No.	Target Variable
1	Hinselmann	**3**	Schiller
2	Cytology	**4**	Biopsy

B. Data Preprocessing

Since the dataset contain missing values, the rows with more than 50% missing values for the attributes are removed. Also the attribute 27 and attribute 28 have missing values for most of the instances; these are also removed from the dataset for further analysis. The missing values in the rest of the dataset are replaced with the mean values of attribute that fit in some specific range for some of the important attributes. The 858 instances are reduced to 757 instances.

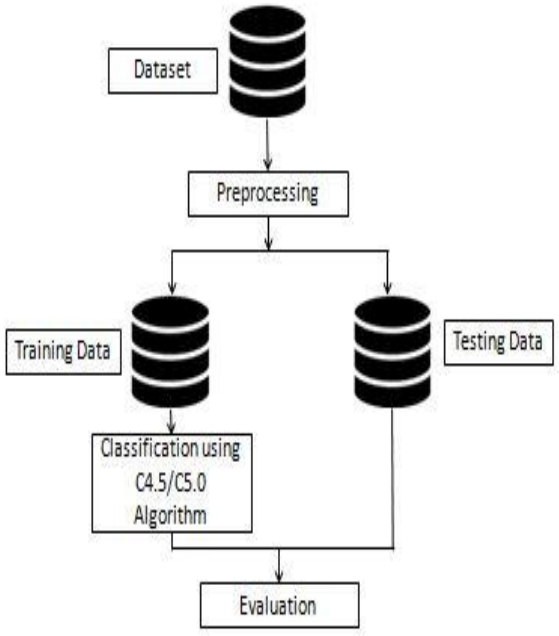

Figure 4: Process 1

C. Dimensionality Reduction

The dimensionality reduction technique is applied in the modified dataset of cervical cancer. The dimensionality is reduced from 36 attributes to 11 attributes. The dimensionality reduction used in this methodology is Principal Component Analysis (PCA).

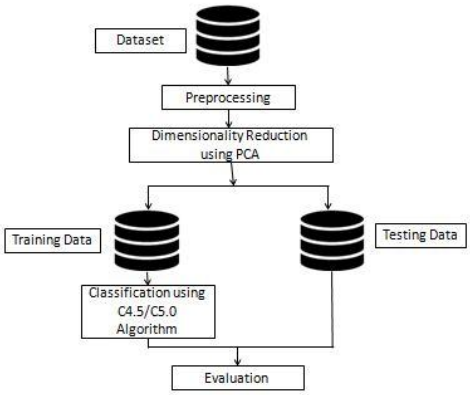

Figure 5: Process 2

a. **Principal Component Analysis (PCA)**

PCA is a linear orthogonal transformation that convert n-dimensional input space to m-dimensional outer space such that m<n and the objective is to minimize the information or variance by discarding (n-m) dimensions. PCA discards the variables that are weakly correlated or have less variance.PCA is feature extraction technique which extracts the highly-correlated components which are the linear mixture of the original features in the dataset known as Principal Components. The first principal component comprises the utmost divergence of the original data after which each consecutive component has the next highest potential variance. These principal components must be orthogonal to each other.

PCA is sensitive to scale of measurement so the variables must be normalized before applying PCA. The algorithm has the following four steps:

i. Standardize the data by subtracting the mean of the dimension across each column to yield a zero mean dataset.

ii. Determine the covariance matrix of the dataset.

$$C^{m\times n} = (c_{ij}, c_{ij} = cov(Dim_i, Dim_j)) \quad (1)$$

$C^{m\times n}$ is a matrix where each entry is the result of calculation of covariance between two separate dimensions.

iii. Calculate eigenvalues and eigenvectors of the covariance matrix.

iv. Reconstruct the data points to a different coordinate system which consists of the l eigenvectors with the maximum eigenvalues.

D. Classification

The rest of the data set is splitted into training and testing dataset in the fraction of 80:20. The classification model is trained on the training dataset and is tested on the undiscovered remaining data.

a. C4.5 algorithm:

C4.5 decision tree algorithm was an improvement to ID3 decision tree algorithms developed by Ross Quinlan. C4.5 need information gain ratio as the splitting principle to determine the attribute. C4.5 overcomes the various shortcomings of the CART and ID3 algorithms like handling missing values, increased accuracy, overcoming over-fitting, and handling both continous and discrete values [15].

b. C5.0 Algorithm

The most recent iteration in the Quinlan's C4.5 algorithm is known as C5.0. The inventor Ross Quinlan claims that it is "several orders of magnitude" faster than C4.5 and many researchers have proved it. Like C4.5 decision tree algorithm, the splitting criterion used by the C5.0 algorithm is gain ratio. Gain ratio is more accurate than the information gain because it is not biased by the multi-valued attributes as it makes decision based on the number and size of the branches of the split while choosing the attribute. Decision tree C5.0 algorithm performs similar to the more advanced machine learning algorithms (Support Vector Machine and Neural Networks) but is easier to understand and implement.

C5.0 algorithm includes all the functionalities of C4.5 algorithm and also includes bunch of new functionalities. These includes (1) generating rulesets from decision trees which are easier to understand [16], (2) efficient handling of the missing and noisy data [16], (3) better pruning algorithm and more efficient in memory and time as compared to C4.5 algorithm [17], (4) provides boosting technique which combines multiple weak classifiers to improve accuracy, (5) supports many new data types, (6) allows to define cost separately for each predicted/actual class pair.

C5.0 builds decision trees in two stages. In the first stage, the tree is grown to fit the data points and in the second phase, tree is pruned by discarding portions that are supposed to have high error rate comparatively. Every subtree is checked to be whether replaced by a leaf or sub-branch and then the performance the whole tree is checked.

c. Splitting Criteria: Gain Ratio

The main drawback of the information gain is that it is biased towards attributes with wide range of values like a unique identifier with a unique value every tuple. The information gain is maximized and is most likely to be chosen as the attribute for splitting. The decision tree algorithm C4.5 and C5.0 uses normalized information gain for splitting the dataset. The gain ratio is defined as

$$Gain\ Ratio\ (X) = \frac{Gain\ (X)}{SplitInfo(X)} \quad (2)$$

Split Information value serves as the inherent knowledge originated by forking the learning record D into r segments, corresponding to r outcomes for attribute X

$$SplitInfo_X(D) = -\sum_{i=1}^{r} \frac{|D_i|}{|D|} \times \log_2 \frac{|D_i|}{|D|} \quad (3)$$

The feature with maximum gain ratio is chosen for splitting which means gain in information is high with this attribute and the partitions are uniform. Low SplitInfo denotes that more tuples are aggregated in few partitions

VI. EXPERIMENTAL RESULT

The resultant dataset is now evaluated using the decision tree C4.5 and C5.0 algorithms. The suggested framework is implemented using the **R programming language version 3.5.3**. The results are calculated original pre-processed dataset and the reduced dataset.

The best approach to weigh the performance of the model is primarily the confusion matrix as shown in Table II. The confusion matrix is a selected table design that indicates the records about real and expected value via a classification version. There are four feasible classification techniques for each occurrence: a true positive (TP), a true negative (TN), a false positive (FP), and a false negative (FN).

Accuracy is computed by the equation (4).

Table 3: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Positive
Actual Negative	False Negative	True Negative

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

Table 4: Results

	Accuracy	Time (in sec)
C4.5	87.64 %	1.813
PCA+C4.5	87.63%	0.923
C5.0	89.8%	1.43
PCA+C5.0	89.26%	0.72

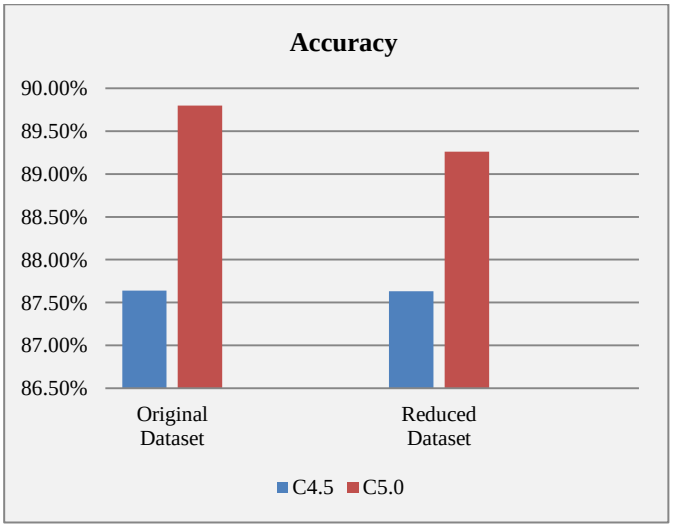

Figure 6: Comparison in Accuracy

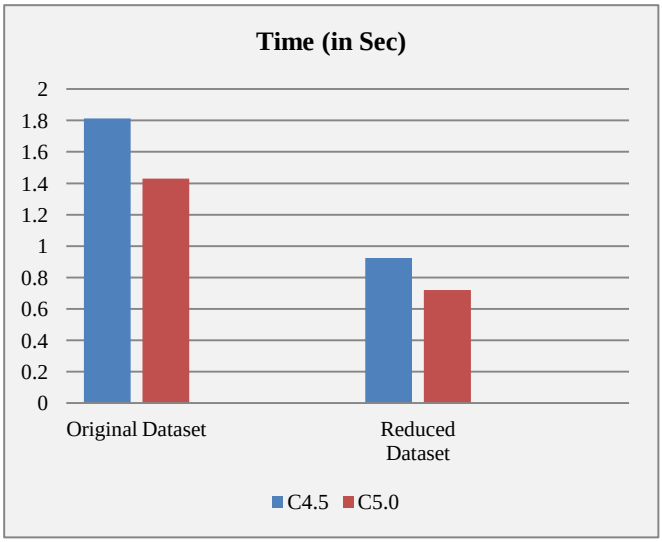

Figure 7: Comparison in Time

VII. CONCLUSION

In the propose methodology, we implemented the C5.0 decision tree algorithm along with Principal Component Analysis (PCA). From the results seen in Fig. 7 there is a significant amount of reduction in time in the reduced dataset as compared to the original dataset. The time of the C4.5 algorithm has reduced from 1.813 seconds of the original dataset to 0.923 seconds of the reduced dataset. Similarly, for the C5.0 algorithm the time has reduced from 1.43 seconds of the original dataset to 0.72 seconds of the reduced dataset. Since time is an important factor while analyzing the large medical data of numerous patients, C5.0 algorithm has improved the time taken. Also C5.0 has performed superior to C4.5 in terms of accuracy as seen in Fig. 6.

REFERENCES

1. W. Wu, H. Zhou, "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches", IEEE Access, Vol. 5, pp. 25189-25195, 2017.
2. American Cancer Society, "Key Statistics for Cervical Cancer", Jan, 2019. Available: https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html.
3. BMC Public Health, "Cervical cancer in Malaysia: can we improve our screening and preventive practice?", Proceedings of the 6th Postgraduate Forum on Health Systems and Policies, Vol. 12, Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-12-S2-A17#Sec3.
4. Y. M. S. Al-Wesabi, Avishek Choudhury, Daehan Won, "Classification of Cervical Cancer Dataset", Proceedings of the 2018 IISE Annual Conference, Orlando, 2018.
5. S. Agarwal, P. Rajan, A. Ujlayan, "Comparative Analysis of Dimensionality Reduction Algorithms, Case Study: PCA", 2017 11th International Conference on Intelligent Systems and Control (ISCO), pp. 255-259, 2017.
6. K. Vijay, K. Selvakumar, "Brain FMRI Clustering Using Interaction K-Means Algorithm with PCA", IEEE ICCSP 2015, pp. 909-913, 2015.
7. D.L. Padmaja, B. Vishnuvardhan, "Comparative Study of Feature Subset Selection Methods for Dimensionality Reduction on Scientific Data", 2016 IEEE 6th International Conference on Advanced Computing, pp. 31-34.
8. L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik, "Dimensionality reduction: A comparative review", Online Preprint, Journal of Machine Learning 2008.
9. Nasution M. Z. F., Sitompul O. S., and Ramli, M. (2017). PCA based feature reduction to improve the accuracy of decision tree c4.5 classification. 2nd International Conference on Computing and Applied Informatics, 1-6.
10. R. Revathy, R. Lawrance, "Comparative Analysis of C4.5 and C5.0 algorithms on crop pest data", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue. 1, pp.50-58, 2017.
11. Prof. Nilima Patil, Prof. Rekha Lathi, and Prof. Vidya Chitre, "Comparison of C5.0 & CART Classification algorithms using pruning technique ", International Journal of Engineering Research & Technology, Vol. 1, Issue 4, June, 2012.
12. Sugandha Agarwal, Dr. Priya Ranjan, Dr. Amit Ujlayan, "Comparative Analysis of Dimensionality Reduction Algorithms, Case Study: PCA", 11th International Conference on Intelligent Systems and Control, pp. 255-259, 2017.
13. P. Douangnoulack and V. Boonjing, "Building Minimal Classification Rules for Breast Cancer Diagnosis," 2018 10th International Conference on Knowledge and Smart Technology (KST), Chiang Mai, 2018, pp. 278-281.
14. https://www.kaggle.com/datasets.
15. D. Ventura, T.R. Martinez, "An Empirical Comparison of Discretization Methods", In Proceedings of the Tenth International Symposium on Computer and Information Sciences, pp. 443-450, 1995.
16. Pandya, R. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. International Journal of Computer Applications, 117 (16), 18–21.
17. Azha, Y. and Afdian, R. (2018). Feature Selection on Pregnancy Risk Classification Using C5.0 Method. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 3(4), 345-350.

AUTHORS PROFILE

Ms. Diksha completed her Masters of Technology in Computer Science and Engineering at IKG Punjab Technical University, Jalandhar, India. She has completed Bachelor in Technology in Computer Science and Engineering from Guru Nanak Dev University, Amritsar, India. Her main research work focuses on Data Mining, Big Data and Machine Learning. She has published a review paper on decision tree based classification algorithms in medical data in leading research journal. She is currently working as an Assistant Professor in Anand College of Engineering and Management, Kapurthala, India. Her area of interest includes supervised learning and has worked on various classification problems.

Dr. Dinesh Gupta is PhD in Computer Science and Engineering from Desh Bhagat University, Punjab, India. He did his M.Tech in IT from Department of CSE, GNDU, Amritsar, India. He has more than 11 years of experience in teaching. Currently he is working as Assistant Professor in department of CSE, IKG Punjab Technical University, India. He has more than 16 publications in leading research journal and has also contributed papers in national and internal conferences. His research areas include machine learning algorithms and cloud computing. He has also authored 4 books related to computer programming and data mining for the undergraduate engineering students. He was also the member of Board of Studies of IKG Punjab Technical University, India.