Performance of Support Vector Machine Kernels (SVM-K) on Breast Cancer (BC) Dataset

Rajesh Kumar Maurya, Sanjay Kumar Yadav, Shwata Agrawal

***Abstract*— Breast cancer (BC) most diagnosed invasive disorder and important cause of casualty for women worldwide. Indian contest BC most commonly spread disease among females. This problem is more alarming to economically developing country like India. Government of India made a lot of effort to make aware the women of the country, but despite of availability of diagnostic tool, prediction of disease in real situation is still a puzzle for researchers. Timely detection and categorization of BC using the evolving techniques like Machine Learning (ML) can show a significant role in BC identification and this could be a preventive policy which effectively reduces the risk of BC patients. Although there are four Kernels in ML, are widely in use but their performance varies with the kind of data available. In this study we, apply four different Kernels such as Linear Kernel (LK), Polynomial Kernel (PK), Sigmoid Kernel (SK) and Radial Basis Function Kernel (RBFK) on BC dataset. We estimated the performance of Support Vector Machine Kernels (SVM-K) on BC dataset .The basic idea is to check the exactness of SVM-K to classify WBCD in terms of effectiveness with respect to accuracy, runtime, specificity and precision. The investigations outcome displays that RBFK provides greater accuracy with minimal errors.**

***Index Terms*: BC Causes, BC Problems, Challenges, ML Techniques, SVM-K, Efficiency, precision, accuracy, run time, specificity, Confusion Matrix.**

I. INTRODUCTION

The first traces of BC are Swelling, abnormal mammogram, dimpling, liposuction, redness, scariness, nipple retraction. Near the beginning most noticeable symptom of BC are changes in discharge, pitting, itching, pain, flaking. Way of life like physical movement and diet and psychological problems, tension, sorrow affect the BC sufferer's life [1] [2]. Indian Medical Research Council (IMRC) started national cancer patent registry program in 1982. IMRC reported that more than 1,300 Indians die daily from BC [3].BC patient counting will double by 2020 in developing nations in year 2018 ICMR estimated that India could record more than 1700000 new cancer cases and more than 800000 deaths by 2020. In 2016, about 1400000 cases of liver cancer recorded [icmr.nic.in].As usual in economically developing nations**,** around 70% of BC cases occur. Good fitness education, awareness program, BC treatment availability effective diagnosis and treatment services will make it good clinical reputation in the country. The timely detection and proper diagnosis of symptoms can make sure long-term patient tolerance from BC [3][4].Before choice of a proper treatment for chronic disease, it is necessary to cautiously study on risky and beneficial reason of all kinds of doctoring [5].ML algorithms are very effective and prevailing tools for categorizing cancerous data set .Number of ML algorithms developed and used to categorize the risky findings. ML techniques are an evolving tool to categories the BC dataset on the basis of the different BC related factors. A Number of studies conducted to find out machine learning application on genetic facts for classification [6]

II. MAJOR RISK FACTOR FOR BC

S.N	**References**	**Risk Factor for BC**	**Descriptions**
A	**www.breastcancercare.org.uk, ww5.komen.org**	**Strong family history of BC patients age factor**	Even if BC might have affected people in a several generations of their family, a trend to be affected at older (>55) ages. A person may also be considered at moderate risk if one close family member developed BC under the age of 40. The BC risk connected to family background may be due to inherited genetic mutations that increase risk. More than 1 first-degree family member (sister, mother or son/daughter) with BC or a female family member diagnosed with BC at a before time

Revised Manuscript Received on July 5, 2019.

Rajesh Kumar Maurya, Department of CS&IT, SHUATS Allahabad, UP, India.

Sanjay Kumar Yadav, Department of CS&IT, SHUATS Allahabad-211007, India-211007

Shwata Agrawal, Department of MCA, ABES Engineering College, NH-24 Ghaziabad,UP,India-201009

B	www.cancer.net	Genetics , Personal History	Genetic testing is recommended for those people who have a familial history of the disease; a biography of BC at any age, and someone close to the family with BC at the age of 50 or less. Out of many BC children showed symptom about 5% are considered to be the result of inherited cancer.
C	www.clinicaltrials.gov, wwwevelandclinic.org	Radiation	Radiotherapy is most commonly used therapeutically in solid tumor/ cancer; however it has its own risk. Breast cancers are highly affected by fluctuation in hormone level, especially at menopausal stage. Social condition also affects the normal hormonal level in women. Altogether these factors act in a synergistic manner.
	www.cdc.gov, www.hsj.gr	Being overweight, pregnancy history	Heaviness may can increase the chances of risk for cancer. Lady has their earliest kid later than age of 30 or not, complete time pregnancies have an elevated danger of BC as compared to women who gave delivery at the age < 30. For reducing the BC danger lactation for 180 Days per kid is important.
D	www.ncbi.nlm.nih.gov, www.breastcancer.org	Menstrual, Drinking	Lady who has begun to adulterate >12 years of age danger of life in BC. The chances of 15% higher danger of BC (women have consumed alcoholic beverages). Lady using smoking (tobacco products) may raise chances of danger of developing BC up to 35%.
E	www.health.harvard.edu, www.breastcancer.org	Plastic, Lack of Exercise Eating unhealthy food	Plastic goods produce certain molecules which are very similar to hormone, thus a probable risk for the increased onset of BC. There are certain reports that suggest that working women's in healthy conditions have less chance of BC. Forecast and safety function in breast preparation is extremely important.

ML to study cancer is an evolving tool for BC data analysis and prediction. There is another huge amount of practical exposures to Data Mining (DM) techniques that use ML methods in the BC region for prediction and prognosis the number of symptoms/cancerous cause. Now ML methods are Widely used to find cancer symptoms using the diagnosis of BC detection by (X-rays). Early finding by doctors requires a consistent opinion and distinguishes between benign tumors and malignant tumors. DNA tests that produce thousands of hereditary estimations used to gather facts from cells and tissue samples on the elements of structural genes, appearance that will be helpful in discovering the pattern [7][8][9]. Forecasting a specific cancer like BC and their response to cure is the most important aim of shrinking tumors, which eventually directs to personal biological therapy. Number of Bioinformatics/computational biology technique developed to forecast, sensitivity based on hereditary patterns. [10].ML tools like SVM, ANN are widely applied in the field of computational biology and are also very effective and successful in BC studies. It seems very difficult to talk to the disease like BC with doctors and research scholar. This is a very broad area, so the patient understands, deciding and psychological adjustment is uncertain. [11][12]. A forecast of repetition of BC is a critical problem for follow-up preparation and successful handling of BC diseases. Various BC computational biological techniques proposed to guess are senselessness based on different substance properties of drugs, while others used the genomic character to find out their consequence. [13]. Modern studies in health prognosis and diagnosis uses the different types of ML categorization algorithms to diagnose the BC disease. To predict disease, the categorization algorithm gives the outcomes as a binary type. When a multilayered biological data set exists, the categorization algorithm reduces the data set to a binary class for simplification using data reduction methods and the algorithm used to predict [14]. Recent studied in medical diagnosis and prognosis uses the number of ML categorization tools to diagnose the BC syndrome. To predict disease, the categorization tools give the outcomes as a binary style. While a multilayered biological BC dataset exists, the categorization tool minimizes the number of dataset to a binary class for generalization using the ML algorithm and fact reduction techniques used to forecast its values [15][16][17].

III.DIFFICULTY IN ANALYSIS OF BC DATA

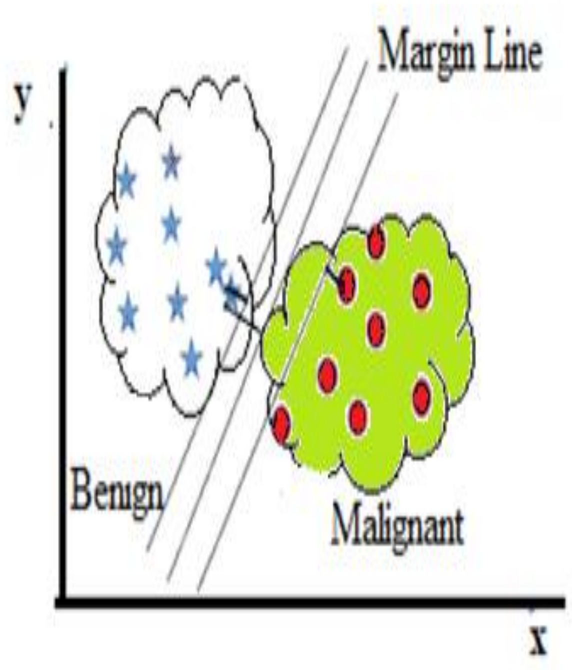
Foremost challenge of ML fields is to design competent and good classifiers for tumor classification [18]. Genomics is new field which initiated a huge inflow of data in biological sciences specially in case of cancer patients [19]. Since there are large number of complex data sets are available thus it need an intricate involvement of computer programmer and biologists to together deal with the tumor classification problem. Pattern of protein expressions and genes are ways to analyze high-transfer data in the form of images photos and are computing tools becoming keys for BC understanding the importance of discovering the remedy and syndrome in the expectations [20].

But, Complexity of the data set induces peculiarity in classification of genes expression using ML. There are mainly following challenges which need to be addressed in genes appearance categorization using ML techniques , is the large amount of gene ,appearance, Investigation of occurrence of error inherent in the dataset ,unique characteristic, classification reliability and accuracy [21][22][23].

IV.SVM FOR BC

ML Classification techniques and Deep Learning (DL) algorithm have huge potential in cancer fact filtering and accomplished of processing capacity of a huge amount of BC facts. ML categorization is a DM task of predicting the value. The main goal of ML filtering is to carefully forecast target group.

A DL network model has become a powerful tool for ML and artificial intelligence [24]. Year 1995 Vladimir Vapnik designed a supervised category ML classification algorithm called SVM. In the field of Bioinformatics, the ML algorithm provides efficient classification tools based on risky pattern reduction. An SVM technique that divides BC data with the greatest margins on two parts – maximum margins between the 2-dimensional planes [25].SVM method used to find 97.7 % accuracy and [26].Using Linear Support Vector Machine (LSVM) find 93.3%, Quadratic Support Vector Machine (QSVM) 92.7%, Fine Gaussian Support Vector Machine (FSVM) 91.3% accuracy investigates the WBCD [27].



V.M

ETHODOLOGY

Using SVM-K the BC patient's records are retrieved from the BC data set using python programming environment .We conduct experiments for evaluating the performances of different kernels of SVM-K in terms of runtime, sensitivity, accuracy, Score(f1), precision, Specificity, confusion_ matrix. Python is a trendy high-level computer programming language developed in 1991.After filtering out of 599 records 550 records selected for this study data of BC patients' set analyzed using libraries of the python in ML run time environment that has an excellent set of functions and methods. Python provides a good integrated development environment (IDE) for implementation of biological observations to build models. For categorization of BC data are used in this analysis, taken from UCI repository [28]. In analyzed dataset had 550 BC patients with tumor ration 66 %(malignant) and 65% (benign).The dataset has 11 attributes ranging from 1-10 shown in the table-1.

Table-1: 11 Features from BC Dataset

Sr. No	Feature	Range
1	Thickness (Clump)	One to Ten
2	Cell Size (Single Epithelial)	One to Ten
3	Bland Chromatin	One to Ten
4	Nuclei (Bare)	One to Ten
5	Nucleoli (Normal)	One to Ten
6	Marginal (Adhesion)	One to Ten
7	Cell Size (Uniformity)	One to Ten
8	Mitoses	One to Ten
9	Cell Shape (Uniformity)	One to Ten
10	ID	Number
11	Class	Zero–One

VI. RESULTS

In Fig.6 SVM-K Performance Comparison BC data examinations of SVM kernels function models are and summarized and compared. .To carry out SVM-K classification and tests them on BC dataset, for evaluating and trains the 4-predictive SVM kernels. Filtered sample of BC data for implementing, n-fold cross-validation techniques are applied using a python environment where n=10.We analyzed BC data as well as find out values in form of competence as well as efficiency.
To test the performance of 4-SVM -K function models at 10-pattern in terms of score (f1), accuracy, precision, runt-time sensitivity. The comparative performance analyses of the results are given in Table 2.

.

SVM-K	Accuracy (%)	Run-time	m-Malignant, b-Benign	Positive predictive value	Sensitivity	Specificity	Score(f1)	Confusion_ matrix		Total
PK	90%	.058	B	87%	100%	71%	93%	100%	0%	66%
			M	100%	71%	100%	83%	29%	71%	34%
LK	94%	.052	B	97%	95%	.95%	96%	94	6%	66%
			M	90%	95%	95%	92%	5%	95%	34%
SK	.96%	.011	B	99%	96%	97%	97%	96%	4%	66%
			M	93%	97%	96%	95%	2%	98%	34%
RBFK	99%	.0014	B	100%	99%	100	99%	98%	2%	64%
			M	97%	100%	99%	99%	0%	100%	34%

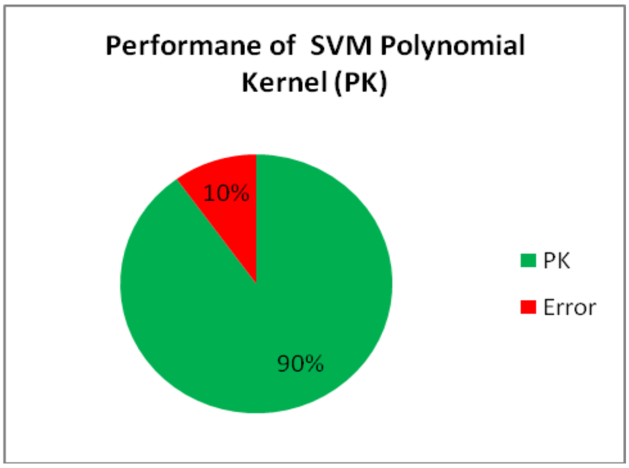


Fig 2: Performance of PK

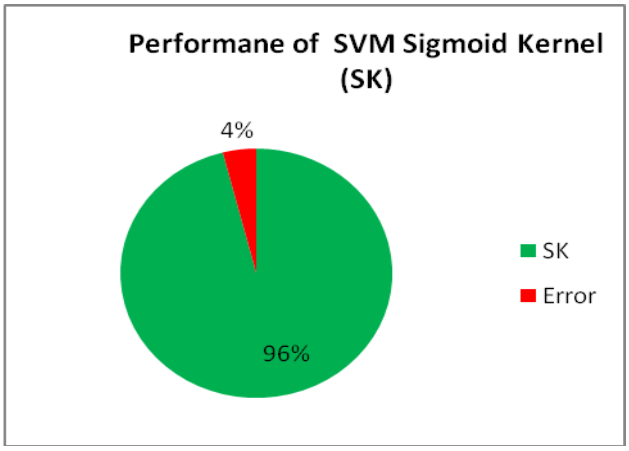


Fig 3: Performance of SK

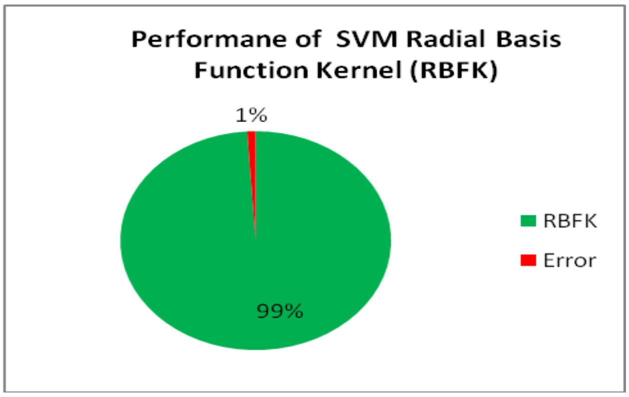


Fig 4: Performance of RBFK

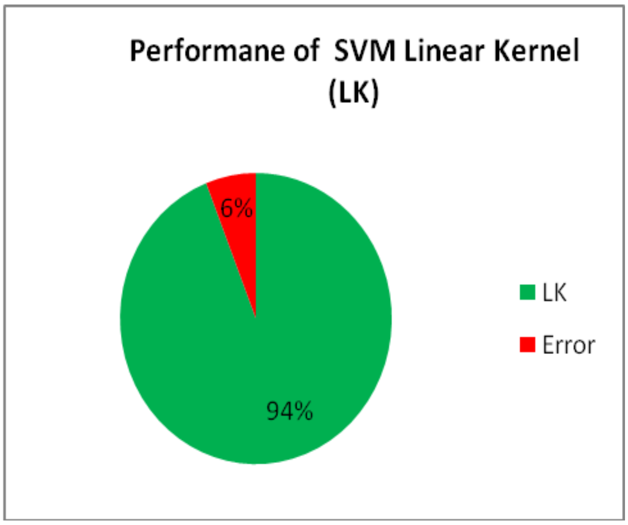


Fig 5: *Performance LK*

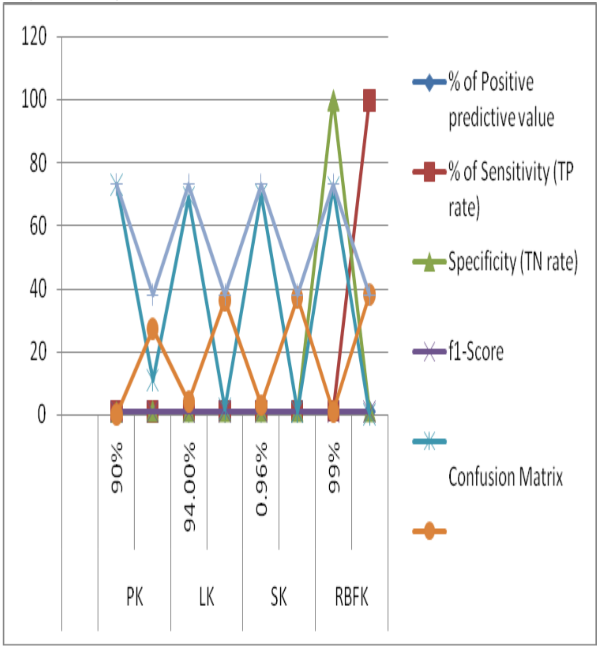


Fig 6: SVM-K performance comparison

VII. CONCLUSION RESULT DISCUSSION

ML is an evolving tool in BC prediction; however different kernel may have a different precision and accuracy of the same data set. To analyze the comparative efficiency and accuracy of the kernels in ML, we conduct experiments on available data set of BC. Data analysis using four different kernels suggests the comparative accuracy of the above system and RBFK is more right on prediction of available data set with a minimum run time. In addition RBFK is able to predict malignancy with 99% outcome which is important for the clinical point of view. This analysis confirms that prediction accuracy of SVM--K classifiers and better categorization and its efficacy in BC prophecy explore a new dimension of making inquiries. This study reflects a way of successful use of SVM and SVM-K techniques to check the different group of BC patients in monitoring the progress of BC at each level. Comparative study and use of different SVM-K classifiers and their BC characteristic choice and research of genetic data assimilation is an excellent approach for doctors and scientists understanding competency in BC prognosis and predictions of diseases. With this small BC data set available RBFK performance fond most accurate, but the analysis needs a large data set to confirm the prediction accuracy. ML tools like SVM-K are seems an evolving tool with better outcomes for BC predictions which could be beneficial to early diagnosis with accuracy in the BC patients and cut the mortality rate in the developing country like India.

REFERENCES

1. Prince, M. J., Wu, F., Guo, Y., Robledo, L. M. G., O'Donnell, M., Sullivan, R., & Yusuf, S. (2015). The burden of disease in older people and implications for health policy and practice. The Lancet, 385(9967), 549-562.
2. Bouchard, C., Blair, S. N., & Haskell, W. L. (2018). Physical activity and health. Human Kinetics.
3. Takiar, R., Nadayil, D., & Nandakumar, A. (2010). Projections of number of cancer cases in India (2010-2020) by cancer groups. Asian Pac J Cancer Prev, 11(4), 1045-9.
4. Unger-Saldaña, K. (2014). Challenges to the early diagnosis and treatment of breast cancer in developing countries. World journal of clinical oncology, 5(3), 465.
5. Gooch, J. C., & Schnabel, F. (2019). Locoregional Recurrence of Breast Cancer. In Clinical Algorithms in General Surgery(pp. 97-100). Springer, Cham.
6. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.
7. Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., ... & Mak, R. H. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: a cancer journal for clinicians.
8. Sahu, B., Mohanty, S. N., & Rout, S. K. (2019). A Hybrid Approach for Breast Cancer Classification and Diagnosis.
9. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, 117693510600200030.
10. Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. British journal of cancer, 108(3), 479.
11. Ing, E., Su, W., Schonlau, M., & Torun, N. (2019). Support Vector Machines and logistic regression to predict temporal artery biopsy outcomes. Canadian journal of ophthalmology. Journal canadien d'ophtalmologie, 54(1), 116-118.
12. Satyananda, V., Ozao-Choy, J., Dauphine, C., & Chen, K. T. (2019). Effect of the Affordable Care Act on Breast Cancer Presentation at a Safety Net Hospital. The American Journal of Surgery.

13. Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(124), 3.
14. Vanaja, S., & Kumar, K. R. (2014). Analysis of feature selection algorithms on classification: a survey. International Journal of Computer Applications, 96(17)
15. Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... & Geessink, O. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), 2199-2210.
16. Gbenga, D. E., Christopher, N., & Yetunde, D. C. (2017). Performance Comparison of Machine Learning Techniques for Breast Cancer Detection. Nova, 6(1), 1-8
17. Huang, S., Cai, N., Pacheco, P. P., Narandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics-Proteomics, 15(1), 41-51.
18. Dubey, U. (2017). Epidemiology of breast cancer in Indian women. Asia-Pacific Journal of Clinical Oncology.
19. Goel, P., & Padole, M. (2019). Bioinformatics: An Application in Information Science. In First International Conference on Artificial Intelligence and Cognitive Computing (pp. 223-238). Springer, Singapore.
20. Dammann, O., & Smart, B. (2019). Health Data Science. In Causation in Population Health Informatics and Data Science(pp. 15-26). Springer, Cham.
21. David, S. K., Saeb, A. T., Rafiullah, M., & Rubeaan, K. (2019). Classification Techniques and Data Mining Tools Used in Medical Bioinformatics. In Big Data Governance and Perspectives in Knowledge Management (pp. 105-126). IGI Global.
22. Shepherd, J., & Perou, C. (2019). Abstract B185: Epithelial cancer cell-expressed genes contribute to clinically relevant immune-based classifications of breast cancer.
23. Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. Information Systems, 28(4), 243-268.
24. Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning-a primer for biologists. arXiv preprint arXiv:1902.04704.
25. Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using.
26. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, 117693510600200030.
27. Karthiga, R., & Narasimhan, K. (2018, March). Automated Diagnosis of Breast Cancer Using Wavelet Based Entropy Features. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 274-279). IEEE.
28. Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]

AUTHORS PROFILE

Rajesh Kumar Maurya M.Tech in CSE,.He works as a Research Scholar in the Department of Computer Science and IT at Sam Higginbottom University of Agriculture, Technology and Sciences Allahabad, India. His areas of interest are Data Mining, Information /Database/ System/Network Security.

Dr. Sanjay Kumar Yadav has received Ph.D. degree in CSE, M.Tech in CSE. He works as an Associate .Professor in the Department of Computer Science and IT, Sam Higginbottom University of Agriculture, Technology and Sciences Allahabad, India. His areas of interest are Distributed Systems, Mobile Ad-hoc Networks, Data Mining,

Shweta Agrawal has received the Master’s degree in Computer Application .She worked as an A.P. in the Department of Computer Application ABES engineering College Ghaziabad, INDIA. Her areas of interest are Data Mining, Software Engineering etc.