ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm

K.Rajesh[1], Dr. Sheila Anand[2]

PG Student[1], Dean (Research) Computer Studies[2]

Rajalakshmi Engineering College, Thandalam

ABSTRACT—***Medical professionals need a reliable prediction methodology to diagnose cancer and distinguish between the different stages in cancer. Classification is a data mining function that assigns items in a collection to target groups or classes. C4.5 classification algorithm has been applied to SEER breast cancer dataset to classify patients into either "Carcinoma in situ" (beginning or pre-cancer stage) or "Malignant potential" group. Pre-processing techniques have been applied to prepare the raw dataset and identify the relevant attributes for classification. Random test samples have been selected from the pre-processed data to obtain classification rules. The rule set obtained was tested with the remaining data. The results are presented and discussed.***

Keywords— **Breast Cancer Diagnosis, Classification, Clinical Data, SEER Dataset, C4.5 Algorithm**

I. INTRODUCTION

Breast cancer occurs due to an uncontrolled growth of cells in the breast tissues [1]. Tumor is an abnormal cell growth that can be either benign or malignant. Benign tumors are non invasive while malignant tumors are cancerous and spread to other parts of the body. Early diagnosis and treatment helps to prevent the spread of cancer. Breast cancer begins in the cells of the lobules or the ducts [2]. 5-10% of cancers are due to an abnormality which is inherited from the parents and about 90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process [3].

According to the statistical reports of WHO, the incidence of breast cancer is the number one form of cancer among women [4]. In the United States (US), approximately one in eight women have a risk of developing breast cancer [5]. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis. Hence, it can be seen from the study that an early diagnosis improves the survival rate. In 2007, it was reported that 202,964 women in the United States were diagnosed with breast cancer and 40,598 women in the United States died because of breast cancer.

A comparison of breast cancer in India with US obtained from Globocon data, shows that the incidence of cancer is 1 in 30 [6]. However, the actual number of cases reported in 2008 were comparable; about 1,82,000 breast cancer cases in the US and 1,15,000 in India. A study at the Cancer Institute, Chennai shows that breast cancer is the second most common cancer among women in Madras and southern India after cervix cancer [7].

Data mining techniques have been extensively applied for breast cancer diagnosis. Diagnosis is used to predict the presence of cancer and differentiate between the malignant and benign cases. In this paper, we have attempted to classify breast cancer data using C4.5 algorithm.

The rest of the paper is organized as follows. In section 2, we briefly discuss related work on application of data mining in breast cancer research. Our work is described in section 3. Section 4 concludes the paper.

II. RELATED WORK

Data Mining is the process of discovering new patterns from large data sets [8]. Classification is a data mining technique based on machine learning which is used to classify each item in a set of data into a set of predefined classes or groups [9]. Classification methods make use of mathematical and statistical techniques such as decision trees, linear programming, neural network and support vector machines.

In this section, we first review a few of the related work on breast cancer diagnosis using data mining techniques. We then discuss some related work on breast cancer analysis of SEER dataset.

Santi Wulan Purnami et al. in their research work used support vector machine for feature selection and classification of breast cancer [10]. They emphasized how 1-norm SVM can be used in feature selection and smooth SVM (SSVM) for classification. Wisconsin breast cancer dataset was used for breast cancer diagnosis. The important attributes were first identified and the diagnosis was carried out based on nine chosen attributes.

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

Farzaneh Keivanfard et al.in their work, have applied feature selection and classification methods based on artificial neural network to classify breast cancer on dynamic Magnetic Resonance Imaging (MRI) [11]. A forward selection method was applied to find the best features for classification. Moreover, artificial neural networks such as Multilayer Preceptron (MLP) neural network, Probabilistic Neural Network (PNN) and Generalized Regression Neural Network (GRNN) were applied to classify breast cancer into two groups; benign and malignant lesions. An accuracy of 100% was achieved using GRNN and PNN. However, specificity obtained in this study cannot be termed accurate because the number of benign cases in the database was not relatively high.

Lambrou et al. introduced a Conformal Predictor based on Genetic Algorithms, and applied to Wisconsin Breast Cancer Diagnosis (WBCD) problem [12]. A rule-based Genetic Algorithms (GAs) was used as a method for building a Conformal Prediction (CP). The resulting algorithm was applied to the problem of breast cancer diagnosis for 683 records without missing values from WDBC dataset. The error rates confirmed the validity of their CP for any given confidence level 1-e, where e is the error rate.

Liu Ya-Qin et al proposed predictive models for breast cancer survivability using SEER data [13]. C5 decision tree algorithm was first used on the imbalanced data and then under sampling was applied to the models to overcome the disadvantage of imbalanced data. Bagging algorithm was then used to increase the performance of the classification for predicting breast cancer survivability. The results obtained showed an accuracy of 0.7678.

Ankit Agrawal et al. in their work analysed the lung cancer data available from the SEER database for developing survival prediction models using data mining techniques [14]. SEER data attributes were classified as demographic attributes, diagnosis attributes, treatment attributes and outcome attributes. Several classification techniques were applied to model the five outcomes of survival after 6 months, 9 months, 1 year, 2 years and 5 years. Attribute selection techniques were applied to identify a small non-redundant set of attributes to develop a prototype mortality risk calculator. It was found that the quality of prediction was retained even with smaller number of non-redundant attributes.

Delen et al, in their work, have developed models for predicting the survivability of diagnosed cases using SEER breast cancer dataset [15]. Two algorithms artificial neural network (ANN) and C5 decision tree were used to develop prediction models. C5 gave an accuracy of 93.6% while ANN gave an accuracy of 91.2%. Bellaachia et al. took the study of Delen et al. as the basis of their research [16]. They have reported that the pre-classification method of Delen et al was not accurate in determining the records of "not survived" class as the cause of death and survivability rate were not taken into consideration. They investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. They have reported that C4.5 algorithm gave the best performance of 86.7% accuracy.

III. PROPOSED METHOD

The processing steps applied to SEER data are given in Figure 1.

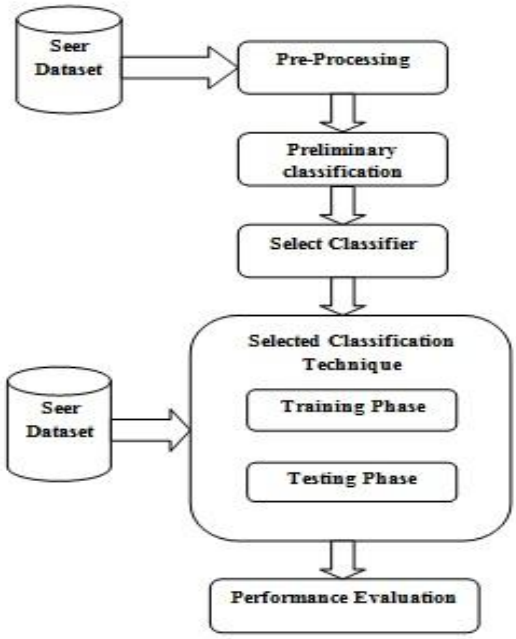

Fig. 1 Processing steps.

SEER (Surveillance, Epidemiology, and End Results) dataset of Program of the National Cancer Institute (NCI) was used for data mining and classification exercise. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28 percent of the US population [17]. SEER database is a premier source for cancer statistics in the United States, which has information on incidence, prevalence and survival from specific geographic areas of the US population as also cancer mortality for the entire country. The dataset used contained data that pertained to all types of cancer cases for the period 1973-2008. Of this,

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

1403 record samples, each having 124 attributes pertained to breast cancer.

A. Pre-processing

Data pre-processing was applied to SEER data to prepare the raw data [8]. Pre-processing is an important step that is used to transform the raw data into a format that makes it possible to apply data mining techniques and also to improve the quality of data [9]. It can be noted from the related work, that attribute selection plays an important role in identifying parameters that are important and significant for proper breast cancer diagnosis. It was also found that the prediction quality was retained even with a small number of non-redundant attributes.

As a first step, non cancer related parameters; also termed as socio demographic parameters were identified and removed. For example, parameters relating to race, ethnicity etc. was discarded. The number of attributes removed in this process was 18 and the total number of attributes was reduced from 124 to 106. Next the attributes having missing values in more than 60% of the records were discarded. For example, the parameter EOD TUMOR SIZE had no values in all the records. 34 attributes were removed in this way and the number of attributes became 72. Then, attributes which were duplicated, that is contained the same values, were overridden or re-coded were discarded. For instance, the attribute HISTOLOGIC TYPE was re-coded as HISTOLOGIC TYPE ICD-O-3. Hence HISTOLOGIC TYPE was discarded. After this process, the number of attributes selected became 56.

The next step was to fill-in actual values for fields which were coded. This was done using the documentation supplied with SEER database. For example, coded data in attribute VITAL STATUS RECODE was replaced with the actual values; code 1 indicated the patient was “alive” and code 2 indicated the patient was “dead”.

The final count of attributes obtained after these processes was 15. Out of this list of 15 attributes, 5 were continuous attributes while others had discrete values. The list of continuous attributes along with descriptions (as given in SEER documentation) is given in Table I.

TABLE I:
SEER CONTINOUS ATTRIBUTES AFTER PRE-PROCESSING

S. No.	Attribute	Description
1	AGE AT DIAGNOSIS	The age of the patient at diagnosis for this cancer which is coded as 1-130 actual age and 999-unknown
2	REGIONAL NODES POSITIVE	Records the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases.
3	SEQUENCE NUMBER—CENTRAL	This sequence number counts all tumors that were reportable in the year they were diagnosed even if the tumors occurred
4	CS TUMOR SIZE	Records the largest dimension or diameter of the primary tumor, and is always recorded in millimeters.
5	CS EXTENSION	Identifies contiguous growth (extension) of the primary tumor within the organ of origin or its direct extension into neighbouring organs.

Finally, the records which had missing values in any of these 5 attributes were discarded. Hence, out of the total 1403 records, 1183 records without missing values were selected for further processing.

B. Preliminary Classification

We next carried out a preliminary analysis on the 1183 data records with different classification techniques using BEHAVIOR CODE ICD-O-3 as the target class and the above mentioned 5 continuous attributes as input attributes. A value of 2 in BEHAVIOR CODE ICD-O-3 denotes “Carcinoma in situ”, while a value of 1 denotes “Malignant Potential” condition. The output rule set obtained for C4.5 algorithm is given in Table II.

TABLE II:
CLASSIFICATION RULES FOR ALL 1183 RECORDS

```
CS EXTENSION < 35.0000
        then BEHAVIOR = Malignant Potential
CS EXTENSION >= 35.0000
        then BEHAVIOR = Carcinoma In Situ
```

It can be seen that the attribute CS EXTENSION was taken as the most relevant when compared to other attributes. If CS EXTENSION is included in the classification attribute set, the other attributes were not even considered. This rule set is obviously not usable as even the attribute CS TUMOR SIZE which indicates the tumor growth, essential to disease classification, is not considered. Hence the attribute CS EXTENSION was discarded.

The classification techniques were applied with the other 4 continuous attributes. The comparison of error rate

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

obtained for the various classification techniques is given in Table III.

TABLE III:
COMPARISON OF CLASSIFICATION TECHNIQUES

S. No	Technique	Error Rate
1	C-RT	0.1014
2	CS-MC4	0.0803
3	C 4.5	0.0761
4	ID3	0.1014
5	K-NN	0.0752
6	LDA	0.1074
7	NAÏVE BAYES	0.1183
8	PLS-LDA	0.1183
9	RND TREE	0.0414
10	SVM	0.1014

It is seen from the table that RND TREE algorithm has the lowest error rate of 0.0414, that is, approximately 4%. The rule set, however was found to be too large and unwieldy and hence becomes difficult to apply to any dataset. Classification Algorithms KNN and C4.5 gives ~92% classification rate. The graphical representation of the results obtained is given in Figure 2.

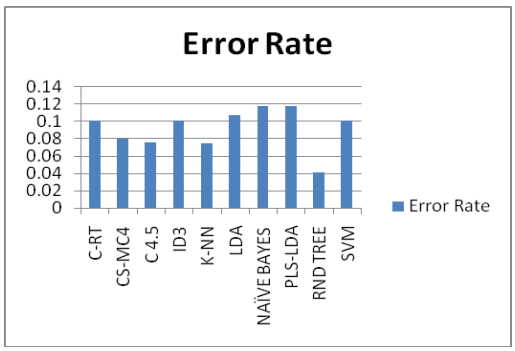

Fig. 2 Comparison of error rates for various classification techniques for 1183 records.

Either of the classification algorithms, KNN or C4.5, could have been chosen to proceed further. We chose C4.5 since it is a well known decision tree induction learning technique that has been successfully and extensively applied for medical data [18].

C. Classification using C4.5

C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the issues not dealt adequately by ID3 [19]. These include avoidance of over fitting the data; reduced error pruning, rule post-pruning, handling continuous attributes and handling data with missing attribute values [20]. C4.5 classification technique uses entropy and information gain for tree splitting. In testing phase we used training data with known result and. C4.5 algorithm was applied to obtain the rule set. In the testing phase, the classification rules obtained were applied to the whole pre-processed data. The results obtained are analysed.

1) Training Phase

We selected three random sets of 500 records from the pre-processed data of 1183 records. This was used as the training data to C4.5 to obtain the classification rule sets. The classification error rate obtained for the three sets of samples is given in Table IV. It can be noted from the table IV that the lowest error rate of 0.599 was obtained for random Sample 2.

TABLE IV:
C4.5 TRAINING PHASE ERROR RATES

Sample	Sample I	Sample II	Sample III
Error rate	0.0640	0.0599	0.0719

As a further verification process, we applied the other classification algorithms to Sample 2 set of 500 records. The graphical representation of the classification results obtained is given in Figure 3.

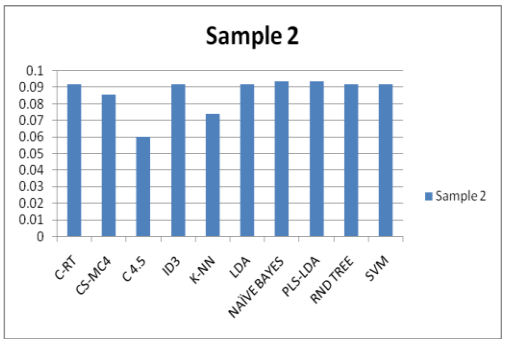

Fig. 3 Comparison of error rates obtained for various classification techniques for Sample 2.

The results obtained are tabulated in Table V.

TABLE V:
COMPARISON OF CLASSIFICATION TECHNIQUES

S. No	Classification Technique	Error rate for Sample set 2
1.	C-RT	0.0918
2.	CS-MC4	0.0858
3.	C 4.5	0.0599
4.	ID3	0.0918
5.	K-NN	0.0739
6.	LDA	0.0918
7.	NAÏVE	0.0938

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

	BAYES	
8.	PLS-LDA	0.0938
9.	RND TREE	0.0918
10.	SVM	0.0918

It is seen from the table and the graph that best results are obtained for C4.5 algorithm. This justifies and validates our choosing C4.5 for classification of SEER data.

2) Testing Phase & Performance Analysis

In testing phase we applied the C4.5 classification rules obtained from random Sample 2 to the complete 1183 records. The actual and predicted values obtained in the classification exercise are shown in the confusion matrix given in Table VI

TABLE VI:
CONFUSION MATRIX FOR SEER

	Malignant Potential	**Carcinoma In Situ**	**Total**
Malignant potential	56 (TP)	64 (FN)	120
Carcinoma In Situ	28 (FP)	1035 (TN)	1063
Total	84	1099	1183

It can be observed from the confusion matrix, that 92 of 1183 records are classified ambiguously. 64 of the "Malignant potential" cases have been classified as "Carcinoma in situ" (false negatives). 28 of the Carcinoma cases have been classified as Malignant (false positives). We now analyse the results obtained using performance measures. Accuracy is the percentage of records correctly classified out of the total records.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{56 + 1035}{56 + 1035 + 28 + 64} = 0.922$$

Sensitivity is the percentage of positive records classified correctly out of all positive records.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} = 0.4666$$

Specificity is the percentage of positive records classified correctly out of all positive records.

$$\text{Specificity} = \frac{TN}{(TN + FP)} = 0.9736$$

The classification rules obtained from the Sample 2 training set in the training phase is given in Table VII.

TABLE VII:
C4.5 CLASSIFICATION RULES FOR SAMPLE SET 2

REGIONAL NODES POSITIVE < 96.5000
CS LYMPH NODES < 25.0000
CS TUMOR SIZE < 47.5000 then BEHAVIOR = Carcinoma In Situ
CS TUMOR SIZE >= 47.5000
AGE AT DIAGNOSIS < 47.5000 then BEHAVIOR = Carcinoma In Situ
AGE AT DIAGNOSIS >= 47.5000
CS TUMOR SIZE < 70.0000 then BEHAVIOR = Malignant Potential
CS TUMOR SIZE >= 70.0000 then BEHAVIOR = Carcinoma In Situ
CS LYMPH NODES >= 25.0000 then BEHAVIOR = Carcinoma In Situ REGIONAL NODES POSITIVE >= 96.5000
CS TUMOR SIZE < 13.5000
CS TUMOR SIZE < 10.5000
CS TUMOR SIZE < 9.5000
AGE AT DIAGNOSIS < 62.5000 then BEHAVIOR = Malignant Potential
AGE AT DIAGNOSIS >= 62.5000
AGE AT DIAGNOSIS < 81.5000 then BEHAVIOR = Carcinoma In Situ
AGE AT DIAGNOSIS >= 81.5000 then BEHAVIOR = Malignant Potential
CS TUMOR SIZE >= 9.5000 then BEHAVIOR = Malignant Potential
CS TUMOR SIZE >= 10.5000 then BEHAVIOR = Carcinoma
CS TUMOR SIZE >= 13.5000 then BEHAVIOR = Carcinoma In Situ

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

IV. CONCLUSION:

We have attempted to classify SEER breast cancer data into the groups of "Carcinoma in situ" and "Malignant potential" using C4.5 algorithm. We used a training set of a random sample of 500 records and then applied the classification rule set obtained to the full breast cancer dataset. We obtained an accuracy of ~94% in the training phase and an accuracy of ~93% in the testing phase. We have compared the performance of C4.5 algorithm with other classification techniques. Future enhancement of this work includes improvisation of the C4.5 algorithms to improve the classification rate to achieve greater accuracy.

ACKNOWLEDGEMENT

This research work is a part of the All India Council for Technical Education (AICTE), India funded Research Promotion Scheme project titled "Efficient Classifier for clinical life data (Parkinson, Breast Cancer and P53 mutants) through feature relevance analysis and classification" with Reference No:8023/RID/RPS-56/2010-11, No:200-62/FIN/04/05/1624. We sincerely acknowledge and thank the National Cancer Institute, USA for permitting us to use the SEER cancer database.

REFERENCES

[1] Breast cancer facts and figures http://www.breastcancer.org/
[2] http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001911
[3] Breast Cancer statistics from Centers for Disease Control and Prevention, http://www.cdc.gov/cancer/breast/statistics/.
[4] D. M. Parkin, F. Bray, J. Ferlay, "Global cancer statistics 2002," CA Cancer J Clin, vol.55, pp. 74-108, 2005.
[5] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).
[6] http://www.breastcancerindia.net/bc/statistics/stat_global.htm
[7] K.Gajalakshmi, V. Shanta, R. Swaminathan, R. Sankaranarayanan, and R. J. Black, "A population-based survival study on female breast cancer in Madras, India", Cancer Institute (WIA), Adyar, Madras, India.
[8] Soumen Chakrabarti, Earl Cox, Eibe Frank, Ralf Hartmut Güting, Jaiwei Han, Xia Jiang, Micheline Kamber, Sam S. Lightstone, Thomas P. Nadeau
Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J. Teorey, Ian H. Witten, "Data Mining-Know it all", Morgan Kaufmann Publishers, 2009
[9] J. Han and M. Kamber , - " Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, 2000.
[10] Santi Wulan Purnami, S.P. Rahayu and Abdullah Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.
[11] Farzaneh Keivanfard , Mohammad Teshnehlab , Mahdi Aliyari Shoorehdeli , "Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging by Using Artificial Neural Networks", Proceedings of the 17th Iranian Conference of Biomedical Engineering (ICBME2010), 3-4 November 2010.
[12] A. Lambrou, H. Papadopoulos, A. Gammerman, "Evolutionary Conformal Prediction for Breast Cancer Diagnosis", Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Larnaca, Cyprus, 5-7 November 2009.
[13] Liu Ya-Qin, Wang Cheng, Zhang Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data ", IEEE 2009.
[14] Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, "A Lung Cancer Mortality Risk Calculator Based on SEER Data", IEEE 2011.
[15] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005.
[16] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.
[17] SEER dataset - Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2008), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2011, based on the November 2010 submission. www.seer.cancer.gov
[18] Knowledge Discovery in Databases. http://www2.cs.uregina.ca/~/c4.5/tutorial.html
[19] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers,1993.
[20] http://en.wikipedia.org /wiki/C4.5_algorithm - C4.5 Algorithm Description

Authors Profile

Dr. Sheila Anand holds a Doctorate in the faculty of Information and Communication Engineering. She has had an illustrious career in industry and academic for nearly 26 years and is presently working as Dean (Research) Computer studies at Rajalakshmi Engineering College. Her other professional qualifications include Certified Information Systems Auditor (CISA), Certified Software Quality Analyst (CSQA) and Certified Information Security Manager (CISM). She is a senior member of IEEE, member in ACM, ISACA, Computer Society of India and IETE. Her current research interests include information security, data mining, computer networks and distributed computing.

Mr. Rajesh. K has completed his B.Tech in Information Technology at Rajiv Gandhi College of Engineering affiliated to Anna University, Chennai, India. Currently he is pursuing his M.E. in Computer Science and Engineering at Rajalakshmi College of Engineering, affiliated to Anna University of Technology, Chennai, India. He is a CISCO certified network associate. His areas of interest include Computer Networks and Data Mining.