Staging Prediction in Cervical Cancer Patients – A Machine Learning Approach

D. Sowjanya Latha[1], P.V. Lakshmi[2] and Sameen Fathima[3]

1. Associate Professor, Department of MCA, AMS School of Informatics, Hyderabad
2. Professor, Head of the Department, Department of Information Technology, GITAM University, Visakhapatnam
3. Professor, Head of the Department, Department of Computer Science, Osmania University, Hyderabad

Abstract: Cervical cancer is one of the most prominent diseases among women worldwide. It is a disease in which cells of the cervix become abnormal and start to grow uncontrollably, forming tumors. Staging specifies the extent to which the cancer has spread from the cervix and to other parts of the body. This is captured from the various diagnoses (tests) like colposcopy, biopsy, imaging studies which are performed after doctor's physical examination. Proper staging is the most important factor in selecting the right treatment plan. Finding out the correct diagnosis and planning of the proper treatment modality plays a crucial role in the field of Oncology. Cervical cancer has a more or less well defined treatment modality. Treatment modalities available are surgery, radiotherapy and chemotherapy or combination of these based on the stage of the cancer. The objective of this paper is to compare the cancer data using data mining algorithms and find out the efficient algorithm that helps in identifying the stage of the cancer. Performance of data mining classification algorithms has been analyzed by computing their accuracy, sensitivity and specificity; and the efficient algorithm has been identified. J48 has outperformed the other algorithms. Accuracy achieved by J48 is 93.03%, specificity and sensitivity are above 0.8 for all the stages of the cancerous data (this accuracy is better than any reported in the literature). The decision tree generated by J48 has picked the attributes that are closely associated with the staging of the cancer. Sensitivity is above 0.7 and specificity for all the stages is above 0.9 using J48. This would be useful for the medical oncologists in identifying the stage of the cancer and plan their treatment accordingly.

Keywords: *Cervical cancer, human papilloma virus, machine aided identification, feature selection, risk factors for cervical cancer, staging, parametria*

1. INTRODUCTION

Cervical cancer is the prominent form of cancer worldwide and ranks as the first common cancer among women in India the age incidence being above 15 years [10]. Worldwide, cervical cancer is the second leading cause of death among women. Around 5, 00, 000 new cases are identified each year and more than 85 percent are in developing countries *(WHO/NCCC)*. This cancer is the only major gynecologic malignancy that is staged clinically according to International Federation of Obstetrics and Gynecology (FIGO) recommendations. Early stage cervical cancer can be cured at an average rate of 80% with either radical surgery or radiation. Accurate cervical cancer staging is crucial for appropriate treatment selection and treatment planning. The prognosis of invasive cervical cancer is based on the stage, size and histology grade of the primary tumor and the size of the lymph nodes. Assessment of disease stage is essential in determining proper management in individual cases. The prognosis of cervical cancer is influenced by local disease extent as determined on the basis of tumor volume, depth and degree of tumor invasion, parametrial invasion, pelvic side wall extension, lymph node involvement, and presence of distant metastases.

The current study is aimed at identifying the most influential risk factors among the other factors in order to reduce the number of deaths and creating awareness among women.

The contents of the paper are organized as follows; in section 2 talks about the biology of the disease and diagnosis is presented. In section 3 we present

an overview of the studies that were made on early diagnosis and staging of the cancer. Section 4 talks about the methods used. Experiments and results are discussed in section 5 followed by conclusion in section 6.

2. CERVICAL CANCER

Cancer of the cervix occurs when the cells of the cervix change in a way that leads to abnormal growth and invasion of other tissues or organs of the body. The normal cells of the cervix first gradually develop pre-cancerous changes that turn into cancer. Since the most common form of cervical cancer starts with pre-cancerous changes, there are two ways to stop this disease from developing. The first way is to find and treat pre-cancers before they become true cancers, and the second is to prevent the pre-cancers. A well-proven way to prevent cervix cancer is to have testing (screening) to find pre-cancers before they can turn into invasive cancer. The Pap test (or Pap smear) and the human papilloma virus (HPV) test are used for this. If a pre-cancer is found it can be treated, stopping cervical cancer before it really starts. The second way is avoiding exposure to HPV. In women, HPV infections occur mainly in younger women and are less common in women older than 30. Certain types of sexual behavior increase a woman's risk of getting genital HPV infection, such as: having sex at an early age, having many sexual partners, having a partner who has had many sex partners, having sex with uncircumcised males[37]. HPV enters the body, usually through a break in the skin, and then infects the cells in the layers of the skin. The virus then replicates or multiplies in the body. The time between first contracting HPV and the appearance of lesions can be weeks to months or even years. HPV is transmitted by skin-to-skin contact. HPV infections that cause skin warts (e.g., plantar or common warts) can be acquired through a cut, but the risk of transmission is low. Walking barefoot in public areas such as the gym or pool can be a risk for infection with the types of HPV that cause plantar warts. HPV infections that cause genital warts are very contagious and are usually contracted through sexual activity with an infected person. A mother with a genital HPV infection may also transmit the virus to the infant during labor. The risk factors for HPV infection include age, number of sexual partners, immune system. There are other risks linked to cervical cancer apart from HPV. These include genetic factors (having close relatives who have suffered from cervical or related cancer), smoking, having a poor diet and or having a weakened immune system (due to surgery, Human Immunodeficiency Virus (HIV) infection, using immunosuppressive drugs). Use of the contraceptive pill (over a long period of time) can also potentially increase the risk of cervical cancer. However, these factors are very minor compared with HPV infection, and are more likely to exert indirect effects; for instance, poor diet can lead to the body not being able to fight off infection so well, so if a person is infected with a high-risk HPV and has a poor diet, this could mean that the body is less well equipped to fight off the virus allowing the virus to infect cells and potentially cause changes that may eventually lead to cancer (these changes are referred to as precancerous changes). The carcinogens in cigarettes can cause damage to the cervical cells, possibly leading to cervical cancer. Studies have shown that smoking can accelerate the cervical damage caused by HPV [40].

Symptoms that may occur in the early stages can include: abnormal vaginal bleeding between periods, after intercourse, or after menopause, continuous vaginal discharge, which may be pale, watery, pink, brown, bloody, or foul-smelling, periods become heavier and last longer than usual. Cervical cancer may spread to the bladder, intestines, lungs, and liver. Patients with cervical cancer do not usually have problems until the cancer is advanced and has spread. Symptoms of advanced cervical cancer may include: back pain, bone pain or fractures, fatigue, leaking of urine or feces from the vagina, leg pain, loss of appetite, pelvic pain, single swollen leg, weight loss[38].

Early cervical cancer does not present any symptoms so regular screening is vital to ensure it is detected as early as possible. Screening prevents 84 out of every 100 cervical cancers that would otherwise develop. Cervical screening involves a sample of cells being taken from the cervix and being examined under a microscope for signs of abnormality. If any abnormal cells are detected then the sample will be graded according to the severity of the abnormalities.

Precancerous changes of the cervix and cervical cancer cannot be seen with the naked eye. Special tests and tools are needed to spot such conditions [12]. Papanicolaou or Pap test is a most common

form of diagnosis for detecting cervical cancer in its early stage. Every woman, above 30 yrs or above 18yrs of age who are sexually active, is recommended to undergo this test every year. If the test shows abnormality then further examination has to be done. If Pap smear results reveal cervical abnormalities, colposcopy is then scheduled. Pieces of tissue are surgically removed (biopsied) during this procedure of colposcopy and sent to a laboratory for examination. Cone biopsy may also be done. A small cone shaped sample of the tissue is removed from the cervix. It is examined under a microscope for signs of cancer[43]. Treatment of cervical cancer depends on: the stage of the cancer, the size and shape of the tumor, the woman's age and general health, desire to have children in the future. Early cervical cancer can be cured by removing or destroying the precancerous or cancerous tissue. There are various surgical ways to do this without removing the uterus or damaging the cervix, so that a woman can still have children in the future. Types of surgery for early cervical cancer include: loop electrosurgical excision procedure (LEEP) -- uses electricity to remove abnormal tissue, Cryotherapy -- freezes abnormal cells, Laser therapy -- uses light to burn abnormal tissue. A hysterectomy (removal of the uterus but not the ovaries) is not often performed for cervical cancer that has not spread. It may be done in women who have repeated LEEP procedures. Treatment for more advanced cervical cancer may include: radical hysterectomy which removes the uterus and much of the surrounding tissues, including lymph nodes and the upper part of the vagina, pelvic exenteration an extreme type of surgery in which all of the organs of the pelvis, including the bladder and rectum, are removed. Radiation may be used to treat cancer that has spread beyond the pelvis, or cancer that has returned. Chemotherapy uses drugs to kill cancer. Pre-cancerous conditions are completely curable when followed up and treated properly [12]. Pre-cancerous conditions may take time to change to cancer. Staging helps in identifying these changes. The objective of this paper is to identify the factor/s in identifying the stage of the cervical cancer so that proper treatment can be given to the patient at the right time.

3. LITERATURE REVIEW

Extensive studies have been made for the early diagnosis, prediction of symptoms, prevention, staging of cervical cancer with reference to demographic factors and clinical data cancerous and non-cancerous patients. Study also made with reference to the cervix images, pap smear images, and genomics and so on. In this section various studies have been quoted referring the research done at various levels.

A classifier system has been designed for cervical cancer diagnosis using bio chemical parameters of cancer patients using Support Vector Machine (SVM) and Classification and Regression Trees (CART) [21]. Unsupervised modeling techniques have been used for feature clustering and classification of cervix images to automatically analyze the uterine cervix images by detecting the detection of cervix boundary and the opening of the cervix [5]. Decision support system for cervical cancer management and staging was designed using soft computing tools like neural networks, genetic algorithm and rough set theory to build an efficient decision making system for pattern classification and rule generation [31]. A new relaxation ranking algorithm was developed to supplement the DNA (Deoxyribonucleic acid) methylation markers in cervical cancer so that the number of validation steps used as part of the experimentation would be reduced for detecting the cancer in cervical scrapings [19]. Demographic data, environmental and genetic factors have been clubbed for analyzing the risk factors for cervical cancer. A model was developed using induction technique in finding out the association among the risk factors and hence generate rules for the management of the disease [29]. A flexible decision based model has been developed using k-means clustering technique for the physician to know the exact conditions for undertaking biopsy test using the demographic data [32]. Multispectral pap smear image classification for cervical cancer detection using a novel SVM-based feature screening method [8]. Identification of risk factors using fuzzy rough sets for detecting cancer at an early stage applied over demographic data [13]. Computerized clinical decision support system for screening cervical cancer by interpreting the free-text pap reports using Natural Language Processing was developed [9]. HPV risk types have been classified using support vector machine (SVM) classifier with gap-spectrum kernel based on k-spectrum method [14]. Diagnostic performance of Magnetic resonance Imaging (MRI) was evaluated

in the pretreatment evaluation of invasive cervical cancer especially for parametrial invasion and lymph node involvement . A study was made on the correlation between MRI involvement and parametrial invasion on histology. It was found that MRI measured tumor volume does not help as a diagnostic criterion rather parametrial invasion is an important factor for cancer treatment because of low accuracy; less than 60% [42]. Medical imaging techniques often detect cancer at its early stage when it is curable and least costly to be treated upon [24].

4. METHODOLOGY

Data pre-processing is essential for successful data mining process [4]. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining [3]. This process reduces the number of features, removes irrelevant, redundant or noisy data thereby improving mining performance such as predictive accuracy and result comprehensibility. The removal of irrelevant and redundant information often improves the performance of the classification algorithms [16]. The feature selection process as shown in fig 1 consists of subset generation, subset evaluation and result validation.

Feature selection algorithms broadly are categorized into a) filter model (univariate, multivariate) b) wrapper model and c) embedded model. Filter model evaluates and selects the subset from the data without involving any data mining algorithm. Wrapper model searches for features that best suites the predetermined mining algorithm; it is computationally expensive than the filter model. Embedded model is a combination of the above two models like C4.5 algorithm.

In our study multivariate filter based model CFS (Correlation based feature selection) and embedded model C4.5 have been applied on the cancer data. The purpose of applying these models is they model the dependencies among the features, which is the basic advantage of feature selection. Cross validation is method for estimating the true error of a model. Cross validation is used to evaluate or compare learning algorithms as follows: in each iteration one or more learning algorithms use k -1 folds of data to learn one or more models, and subsequently the learned models are asked to make predictions about the data in the validation fold. The performance of each learning algorithm on each fold can be tracked using some predetermined performance metric like accuracy. In our study we applied ten-fold cross validation.

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. This classification is based on the type of classification like rule based, tree based, function based, fuzzy rule based and so on.

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. This technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g., information gain, Gini index, and Chi-squared test) to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. In our study we have used J48 (implementation of C4.5 in Weka). C4.5 [34] is a software extension of the basic ID3 algorithm designed by Quinlan.

C4.5 is a supervised learning algorithm. The algorithm analyzes the training set and builds a classifier that must be able to correctly classify both training and test examples. The classifier used by C4.5 is a decision tree.

Algorithm: C4.5

1. Check for the base case
2. Find the attribute with highest information gain (A-best)
3. Partition S into S_1, S_2, .. according to the values of A-best
4. Repeat the steps for S_1, S_2, ..

The base cases are as follows:

- All the examples from the training set belong to the same class
- The training set is empty
- The attribute list is empty

Rule-based expert systems are often applied to classification problems in various application fields, like fault detection, biology, and medicine. Fuzzy logic can improve such classification and decision support systems by using fuzzy sets to define overlapping class definitions. The application of fuzzy if-then rules also improves the interpretability of the results and provides more insight into the classifier structure and decision making process. In our study we have applied NN and Fuzzy RoughNN [28] as these are applied in the literature.

Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perception and can distinguish data that is not linearly separable.

Bayesian classifiers assign the most likely class to a given example described by its feature vector. Learning such classifiers can be greatly simplified by assuming that features are independent given class, that is, $P(X|C) = \Pi P(X_i | C)$, where $X = (X_1, X_2, \ldots X_n)$ is a feature vector and C is a class. Despite this unrealistic assumption, the resulting classifier known as N*aive Bayes* is remarkably successful in practice, often competing with much more sophisticated techniques. Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management.

5. EXPERIMENTS AND RESULTS

The dataset consists of 203 cervical cancer patient cases *(Courtesy: Mitra et al., 2000)*. It consists of 21 boolean features that indicate the signs and symptoms observed upon physical examination containing the 4 stages of the cancer. The 21 Boolean input features refer to *Vulva: healthy (Vu(h)), Vulva: lesioned (Vu(l))* , *Vagina: healthy (Va(h)), Vagina:spread to upper part (Va(u)), Vagina: spread to middle part (Va(m)), Vagina: spread to lower part (Va(l)), Cervix:healthy (Cx(h)), Cervix: eroded (Cx(e)), Cervix: small ulcer (Cx(su)), Cervix: ulcerative growth (Cx(u)), Cervix: proliferative growth (Cx(p)), Cervix: ulcero-proliferative growth (Cx(l)), Paracervix: free (PCx(f)), Paracervix: infiltrated (PCx(i)), Urinary bladder base: soft (BB(s)), Urinary bladder base: hard (BB(h)), Rectrovaginal septum: free (RVS(f)), Rectrovaginal septum: infiltrated (RVS(i)), Parametrium: free (Para(f)), Parametrium: spread, but not upto (Para(nu))* and *Parametrium: spread upto (Para(u))*, respectively. Staging of cervical cancer is given by FIGO.

The purpose of this study is to identify the attributes and extract rules for easily identifying the stage based on the signs and symptoms to identify the right treatment.

Accuracy of C4.5 is better as compared to the accuracy achieved in [31]. Accuracies achieved by Mitra are 81.5 and 80.2 on training and test data respectively whereas the accuracies obtained in this paper are 93.06 and 90.19 for training and test data respectively after applying Correlation based Feature Selection (Cfs). Among the algorithms compared C4.5 (J48 is its implementation in Weka) has outperformed as seen in table 1.

Precision is a measure of the accuracy provided that a specific class has been predicted. Recall or Sensitivity is a measure of the ability of a prediction model to select instances of a certain class from a dataset. It corresponds to the true positive rate. Specificity is a measure commonly used in two class problems where one is more interested in a particular class. It corresponds to the true-negative rate.

Sensitivity is computed using the following formula (TP/(TP+FN)) where TP is True Positive, FN is False Negative. Specificity is computed using the following formula (TN/(FP+TN)) where TN is True Negative, FP is False Positive, TN is True Negative.

In this paper sensitivity and specificity values have been computed for all these algorithms as shown in table 2. J48 has values greater than 0.9 for both sensitivity and specificity. The results show that J48 is able to diagnose all the stages of the cancer.

The results shown in table 3 project the better performance of C4.5 algorithm with feature selection in the working paper compared to the other similar studies. C4.5 algorithm is efficient, the rules generated are easier to understand. A major disadvantage that has been identified in this paper is that the rules generated by C4.5 are clear and precise but are not sufficient enough for predicting the stages of the cancer. A further study needs to be made to enhance the performance of the algorithm. This would be the scope of this paper.

The rules generated by J48 algorithm are as follows:

If {Para(u)=Y and BB(s)}=N then stage = IV

If {Para(u) = Y and BB(s)}=Y then stage = III

If {Para(u)= N and (Para(f)= N) or

(Para(f) = Y and Va(u) = Y)} then stage = II

If {Para(u) = Y and Va(u) = N} then stage = I

The decision tree generated by J48 algorithm is shown in figure 2. The decision tree and the rules generated give a picture of the most influential attributes for identifying the stage of the cancer.

6. CONCLUSION

The comparative study of multiple classifiers identifying the stage of cervical cancer using a dataset of size 221 records provided us with an insight into the predictive ability of different data mining methods. Accuracy achieved by J48 algorithm is better than any given in the literature. Sensitivity and specificity analysis on these algorithms provided us with the prioritized importance of the prognostic factors that lead to the staging of the cancer. This analysis was not performed in any given in the literature. Data analysis was done using 10-fold cross validation. We can conclude saying that by applying data mining algorithms the invaluable efforts of the medical professionals can be enhanced to save more human lives by giving proper treatment at the right time.

ACKNOWLEDGMENTS

We would like to thank Dr Kodati Vijaya Lakshmi, Vasavi Hospital, Dr Sunil and Dr Kameswari, Fernandez Hospital, Dr Benjamin and Dr Usha, MNJC Hospital, Dr Sudarshan and Dr Sushma, Indo-American Cancer Hospital who have helped in gathering information about the disease and to interact with a few patients. Special thanks to Dr Sushmita Mitra and Prof Pabitra Mitra, Indian Statistical Institute, Calcutta for sharing the data.

REFERENCES:

Ambika Satija, *Cervical Cancer in India*, South Asia Centre for Chronic Disease.

American College of Obstetricians and Gynecologists. ACOG Practice Bulletin No. 99: *Management of abnormal cervical cytology and histology*. *Obstet Gynecol*. 2008;112:1419-1444.

H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*, eds. Boston: Kluwer Academic, 1998, second printing, 2001.

Huan Liu, Lei Yu, *Toward Integrating Feature Selection Algorithms for Classification and*

Clustering, IEEE Transactions on Knowledge and Data Engineering, Vol 17, No 4, April 2005

Hayit Greenspan etal., *Automatic Detection of Anatomical Landmarks in Uterine Cervix Images*, IEEE Transactions on Medical Imaging, Vol 28, No. 3, March 200

J Han, M Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Second Edition, 2006

Jyotsna A Saonere, *Awareness screening programme reduces the risk of cervical cancer in women*, African Journal of Pharmacy and Pharmacology, Vol. 4(6), pp 314-323, June 2010

Jiayong Zhang, Yanxi Liu, *Cervical Cancer detection using SVM based Feature Screening*, Carnegie Mellon University, NIH program

Kavishvar, Kathy, Michael et. al., *Clinical Decision support with automated text processing for cervical cancer screening*, Biomedical Journal, 2012. Available online at www.jamia.org/content/early/recent.

Kaarthigeyan K, *Cervical Cancer in India and HPV Vacination*,Indian J Med Paediatr Oncol 2012, 33:7-12

Katz VL, Lentz GM, Lobo RA, Gershenson DM,, Noller KL. *Intraepithelial neoplasia of the lower genital tract (cervix, vulva): Etiology, screening, diagnostic techniques, management.* In: eds. *Comprehensive Gynecology*. 5th ed. Philadelphia, Pa: Mosby Elsevier; 2007:chap 28.

Krishnakumar Duraisamy et al., *Methods of Detecting Cervical Cancer*, Advances in Biological Research 5(4), 2011, pp 226-232, ISSN 1992-0067

Kuzhali, Rajendran, et al, *Feature Selection Algorithm Using Fuzzy Rough Sets for Predicting Cervical Cancer Risks*, Modern Applied Science, Vol 4 No 8, August 2010

Kiam and Zhang, *HPV Risk Type Classification from Protein Sequences using SVM*, LCNS 3907, pp 57-66, 2006

Kurkure AP, and Yeole BB, *Social inequalities in cancer with special reference to South Asian countries*, *Asian Pacific Journal of Cancer Prevention*, 7(1) (Jan-March 2006): 36-40.

Lei Yu, Huan Liu, *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, Proceedings of twentieth International Conference on Machine Learning (ICML), 2003

Margaret H Dunham, S Sridhar, *Data Mining: Introductory and Advanced Topics*, Pearson Education, 2006

Martin-Hirsch PPL, Paraskevaidis E, Bryant A, Dickinson HO, Keep SL. *Surgery for cervical intraepithelial neoplasia*. *Cochrane Database of Systematic Reviews* 2010, Issue 6. Art. No.: CD001318.

Mate Ongenaert et al., *Discovery of DNA methylation markers in cervical cancer using relaxation ranking*, BMC Medical Genomics, November 2008

Mohd Otham, Thomas Yau, *Comparision of different Classification Techniques using WEKA for Breast Cancer*, Biomed 06, IFMBE Proceedings 15, pp 520-523, 2007

Nester Jeyakumar M et al., *Improved Classifier performance through genetic algorithm for cervical cancer prediction*, Journal of Research in Bioinformatics, Apr 2012.

Pavani Sowjanya et al., *Prevalence and distribution of high-risk human papilloma virus (HPV) types in invasive squamous cell carcinoma of the cervix and in normal women in Andhra Pradesh, India*, BMC Infectious Diseases 2005, 5: 116

Pieman, Wahid, Azuraliza, *A Comparative study for various methods of classification*, 2012 International conference on Information and Computer Networks, IPCSIT vol. 27(2012)

Polidais LLC, *Medical Imaging in Cancer care: Charting the progress Executive Summary*, New England Journal of Medicine, 342.1, 2000, pg 42-49

Rakesh Dikshit, et. al, *Cancer mortality in India: a nationally representative survey*, [Document

online], Available from http://www.thelancet.com, March 2012, DOI:10.1016/S0140-6736(12)60358-4

R Prasad, *Cancer Killed 5.56 lakh in India in 2010*, The Hindu, Chennai, March 28, 2012.

Ross Quinlan J., *Machine Learning*, 1st ed. Morgan Kaufmann Publishers Inc., 1993

Bustanur Rosidi, et.al., *Calssification of Cervical Cells Based on Labeled Colour Intensity Distribution*, International Journal of Biology and Biomedical Engineering, 2011

Seung Hee Hoo, Sun Ha Jee, Jong Eun, Jong Sup, *Analysis on risk factors for cervical cancer using Induction technique*, Expert Systems with applications, Elsevier, 2004

Shravya Reddy Konda, *A Comparative evaluation of Symbolic Learning Methods and Neural Learning Methods*, University of Maryland

Sushmita Mitra, Pabitra Mitra, *Staging of Cervical Cancer with Soft Computing*, IEEE Transactions on Biomedical Engineering, Vol 47, No 7, July 2000

Thangavel, Jagannathan, Easmi, *Data Mining Approach to Cervical Cancer Patients Analysis using Clustering Technique*, Asian Journal of Information Technology, 5(4), 2006

Thales Senh Korting, *C4.5 algorithm and multivariate decision trees*, Image Processing Division, National Institute for Space Research, Brazil

Veronica S Moertinin, *Towards the use of C4.5 algorithm for classifying banking dataset*, Integral Vol 8 No 2, October 2003

Zhang Liu, Tong Zhao, Yanxi Liu, *SVM Based Feature Screening applied to Hierarchical Cervical Cancer Detection*, Carnegie Mellon University, NIH program

WEKA (Data Mining Software). Available at http://www.cs.waikato.ac.nz/ml/weka/. 2006.

Cervical cancer incidence [Document-online] Available from http://www.medindia.net/patients/patientinfo/cervicalcancer-incidence.htm

Cervical Cancer Risk Factors [Document-online] Available from http://www.cancer.org/Cancer/CervicalCancer/DetailedGuide/cervical-cancer-risk-factors

A Jena, et.al., *Parametrial invasion in carcinoma of cervix: role of MRI measured tumour volume*, The British Journal of Radiology, 78 (2005), 1075 – 1077

http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004359

Authors

Ms D Sowjanya Latha is an Associate Professor at AMS School of Informatics, Hyderabad, India. She is MCA with 14 years of experience. She has 3 publications to her credit. She is currently pursuing her PhD (CSE) from GITAM University. Her areas of research are Data Mining, Bioinformatics, Network Security.

Dr PV Lakshmi is Professor & Head, Dept of IT, GITAM University, Visakhapatnam, India. She is PhD (CSE) and MTech (CSE) from Andhra University, Visakhapatnam, India. She has around 14 papers published in International Journals, 2 papers in National journals, 8 in International conferences, 2 in National conferences. She has been awarded as the *Best Academician* in 2010 from CSE Dept, GITAM University. She has around 16 years of teaching experience. Her areas of research are Bioinformatics, Cryptography, Network Security.

Dr Sameen Fatima is Professor & Head, Dept of CSE, Osmania University, Hyderabad, India. She is PhD (CSE) from Osmania University, Hyderabad, India, MS (CSE) from University of Massacheusetts, USA, MPhil (Computer Methods), University of Hyderabad, She has around 14 publications to her credit. Her areas of research are Information Retrieval System, Artificial Intelligence, Machine Learning.

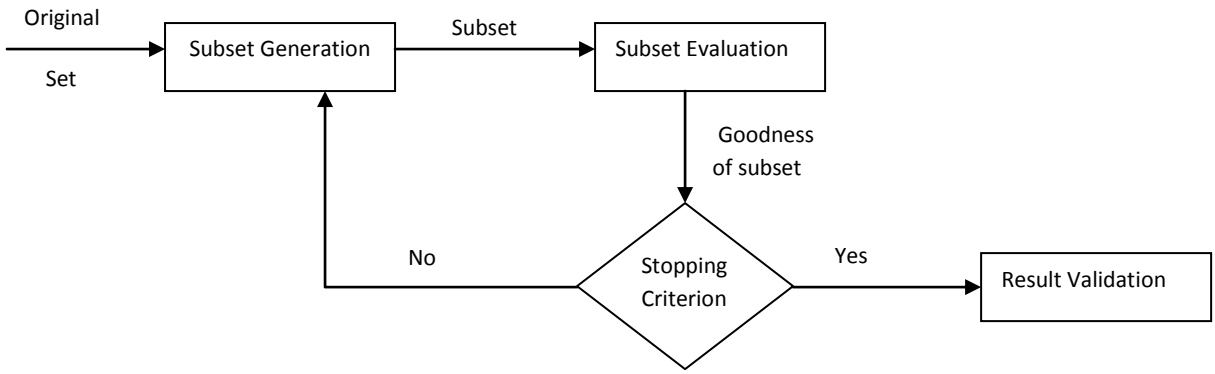

Figure 1: Unified process of feature selection

Table 1: Performance analysis using accuracy on data mining algorithms

	Accuracy				
	Train Set	*Test Set*	*Full Set (10-fold cross validation with CFS)*	*Full Set (10-fold cross validation without CFS)*	*10-fold cross validation with Wrapper and Subset Evaluator*
J48	*93.069*	*90.196*	*87.192*	*87.192*	*87.192*
SVM	*86.418*	*92.156*	*80.295*	*81.773*	*85.221*
FuzzyRoughNN	*92.079*	*100*	*61.0837*	*66.5025*	*67.487*
NN	*89.108*	*98.039*	*80.295*	*84.236*	*85.714*
NaiveBayes	*93.069*	*92.156*	*85.221*	*84.729*	*87.684*
MLP	*95.049*	*94.117*	*82.758*	*78.325*	*87.192*

Table 2(A): Comparison of the Performance of mining algorithms

	Sensitivity						*Specificity*					
	J48	*SVM*	*Fuzzy RoughNN*	*NN*	*NaïveBayes*	*MLP*	*J48*	*SVM*	*Fuzzy RoughNN*	*NN*	*NaïveBayes*	*MLP*
Stage I	*0.857*	*0.143*	*1*	*0.5*	*0.571*	*0.571*	*0.968*	*0.983*	*0.893*	*0.971*	*0.973*	*0.968*
Stage II	*0.710*	*0.789*	*0.421*	*0.632*	*0.737*	*0.711*	*0.960*	*0.9*	*0.775*	*0.924*	*0.940*	*0.935*
Stage III	*0.926*	*0.941*	*0.638*	*0.971*	*0.956*	*0.926*	*0.922*	*0.776*	*0.857*	*0.671*	*0.835*	*0.849*
Stage IV	*0.8*	*0.2*	*0.066*	*0*	*0.467*	*0.467*	*0.963*	*0.973*	*0.990*	*0.994*	*0.978*	*0.962*

Table 2(B): Comparison of the Performance of mining algorithms

	AUC					
	J48	*SVM*	*Fuzzy RoughNN*	*NN*	*NaïveBayes*	*MLP*
Stage I	*0.95*	*0.56*	*0.95*	*0.96*	*0.96*	*0.95*
Stage II	*0.86*	*0.84*	*0.74*	*0.86*	*0.88*	*0.85*
Stage III	*0.90*	*0.85*	*0.90*	*0.93*	*0.92*	*0.91*
Stage IV	*0.89*	*0.58*	*0.87*	*0.87*	*0.81*	*0.87*

Table 3: Comparison with other similar studies

	Methods Used	*Sensitivity*	*Specificity*	*Accuracy*
[31] Pabitra, 2001 (Staging data, clinical)	*RoughSet Theory* *ID3* *C4.5*			*81.03* *82.74* *80.2*
[42] Jena, 2005 (MR images of cervical cancer)		*59.26*	*61.54*	*60.95*
[29] Seung Hee Ho, et.al., 2004 (demographic, environmental and genetic factors)	*CHAID* *Logostic Regression*	*64.00* *40.80*	*77.83* *88.70*	*72.96* *71.83*
[28] Bustanur Rosidi, et.al., 2011 (Cervix cells)	*Labeled Color Intensity Distribution*	*78.6*	*75.9*	*77.0*
Working Paper (Staging data, clinical)	*C4.5 (with feature selection)* *SVM* *FuzzyRoughNN* *NN* *NaiveBayes* *MLP*	*82* *51* *53* *52* *68* *67*	*95* *91* *88* *89* *93* *93*	*87.19* *80.29* *61.08* *80.29* *85.22* *82.75*