CHAPTER IV

STAGING PREDICTION IN CERVICAL CANCER PATIENTS – A MACHINE LEARNING APPROACH

4.1 INTRODUCTION

Cervical cancer is the most prominent form of cancer worldwide and ranks as the first common cancer among the women in India the age incidence being above 15 years [10]. Worldwide, cervical cancer has been identified as 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 according to the statistics given by World Health Organization (*WHO)* and National Cervical Cancer Coalition (*NCCC)*. This cancer is the only major gynaecologic malignancy that is staged clinically according to International Federation of Obstetrics and Gynaecology (FIGO) recommendations.

Radical surgery or radiation can be used to cure cervical cancer at a very early stage at an average rate of 80 percent with either radical surgery or radiation. Identifying the stage of the cancer accurately is crucial for appropriate treatment selection and treatment planning. Stage, size of the tumor, histology grading of the primary tumor, size of the lymph nodes are important factors for the prognosis of invasive cancer.

Assessing the stage of the disease is crucial for the proper management of the individual cases. The extent to which cervical cancer has spread is dependent upon the invasion of the disease locally. This is determined by certain factors like the volume of the tumor, depth and degree of the invasion of the tumor, parametrial invasion, lymph node involvement, pelvic side wall extension etc. The aim of the research is to identify the most influential risk factors among the other factors in order to reduce the number of deaths and creating awareness among women.

4.2 STAGING OF CERVICAL CANCER

A thorough understanding of the spread of the cancer will be very much useful to the doctors as well as the medical practitioners in identifying the seriousness and the kind of treatment to be provided. The sooner the prediction of the stage; the better would be the nature of the treatment. Staging prediction is another serious area of research as this reduces the risk of cancer deaths.

There are several methods to stop the pre-cancerous changes that can turn into cancer. Find and treat the pre-cancers before the virus spreads, prevent the pre-cancers from gaining entry into the body. Pap test or Pap smear is one way of identifying the pre-cancer. Women above 30 years need to undergo this test once in a year. The pre-cancer can be treated once identified and it is not risky to stop the cancer from spreading. Preventing the exposure of HPV among women is another way of curbing the pre-cancer.

Women below 30 years are easily prone to be infected with this virus. HPV gains entry into the body through skin-skin contact. It may take weeks or months or even years to identify the virus lesion after being contracted with the virus (HPV). Hence it is very difficult to identify the virus; prevention is the only better way of treating this virus. Concerned to this issue several studies have been made in order to identify the reasons behind the contract of the disease. There is a need for creating awareness among women.

Regular screening to identify the risk of the cancer at an early stage is very important as cervical cancer does not present any symptoms. 84 percent of the cervical cancers can be prevented through screening. A small sample of cells are removed from the cervix, examined using a microscope to find out any signs of abnormality. Upon the detection of any abnormality in the cells the sample is graded based on the degree of severity of the abnormalities.

Symptoms of the cancer cannot be perceived until the cancer has advanced and spread. Pre-cancerous can be cured completely if treated properly at the right time and followed. 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.

4.3 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). 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.

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. 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.

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.

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. Multispectral pap smear image classification for cervical cancer detection using a novel SVM-based feature screening method. Identification of risk factors using fuzzy rough sets for detecting cancer at an early stage applied over demographic data.

Computerized clinical decision support system for screening cervical cancer by interpreting the free-text pap reports using Natural Language Processing was developed. HPV risk types have been classified using SVM classifier with gap-spectrum kernel based on k-spectrum method.

Diagnostic performance of Magnetic resonance Imaging (MRI) was evaluated in the pre-treatment evaluation of invasive cervical cancer specifically 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 percent.

Medical imaging techniques often detect cancer at its early stage when it is curable and least costly to be treated upon. Several studies have been made in this direction of identifying the stage of the cancer in order to predict the right treatment for the longevity of the patient.

4.4 METHODOLOGY

Data pre-processing is essential for successful data mining process. Feature selection is one of the important and frequently used techniques in data pre-processing for data mining. This process reduces the number of features, removes irrelevant, redundant or noisy data thereby improving the performance of data mining through predictive accuracy and result comprehensibility.

The removal of irrelevant and redundant information often improves the performance of the classification algorithms. The feature selection process as shown in figure4.1 consists of subset generation, subset evaluation and result validation.

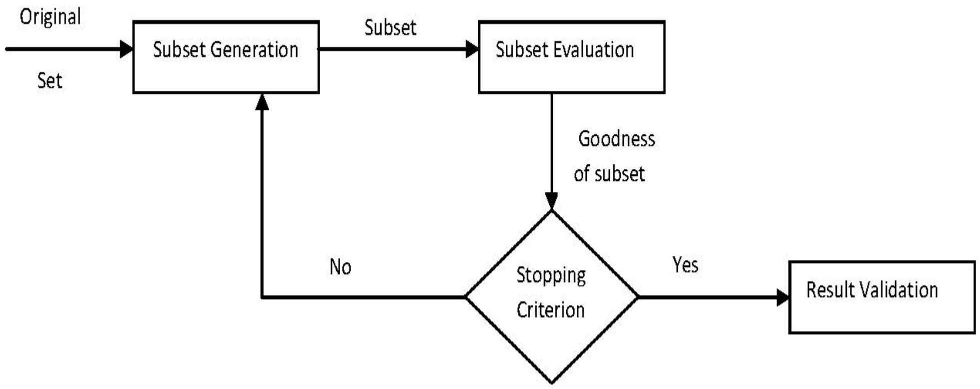

Figure 4.1: Unified Process of feature Selection

Figure 4.1 presents the general process used for selecting a subset from a given set of features. Initially the feature set is given as input and using filter based approaches a subset is generated. These filter based approaches filter the features based on several factors like Attribute Selector, discretize process, based on the class order, based on the classification order etc.

Mostly attribute selection filter is used which is a supervised filter based approach. Once the feature subset is extracted this is evaluated for its goodness. This process is repeated till a stopping criterion is met. If the evaluator yields a good result then the features with the goodness values are picked up as the resultant set. The remaining features undergo the process till the goodness factor is enhanced.

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, a method used for evaluating or comparing the learning algorithms works as follows: for each iteration the learning algorithm uses k -1 folds of data to learn one or more models. Subsequently the learned models make predictions about the data in the validation fold. The performance of the learning algorithm in every fold can be tracked using certain predetermined performance metrics like accuracy. In this research ten-fold cross validation has been applied for evaluating and comparing the algorithms.

Systems with classifier construction are one of the most commonly used tools in data mining. Such systems take a collection of cases as input. Each case belongs to a small number of classes that is described by its values for a fixed set of attributes. Classifier is generated as an output that can predict the class of a new case accurately. 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 used popularly in the field of information systems for mining data. This technique splits the data into branches recursively to construct a tree in order to improve the prediction accuracy. To enhance the prediction accuracy mathematical algorithms like information gain, Gini index, and Chi-squared test are used. The purpose of these algorithms is to identify a variable and corresponding threshold for the variable that splits the input into two or more subgroups. This process is repeated at every leaf node until the complete tree is constructed.

The objective of the splitting algorithm is in finding a variable-threshold pair that would maximize the order of the resulting two or more subgroups of samples. In our study J48 has been used. 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 based on decision tree based approach.

Algorithm: C4.5

1. Check for the base case

2. Find the attribute with highest information gain (A-best)

3. Partition S into S1, S2, .. Sn according to the values of A-best

4. Repeat the steps for S1, S2, ..

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

Fault detection, biology, and medicine fields use rule based expert systems for classifying the problems. Fuzzy logic will enhance the classification of such systems; decision support systems with fuzzy sets will identify the overlaps in the class definitions. Fuzzy if..then rules give a clear understanding of the results and give more insight into the structure of the classifier and the decision making process. NN and Fuzzy RoughNN have been applied in the current study.

SVM's are supervised learning models with the respective learning algorithms that help in analyzing the data, recognizing patterns used in classification and regression analysis. The most basic form of an SVM is used as a non-probablistic binary linear classifier that classifies the given input into their respective binary classes.

SVM training algorithm builds a model with the set of training examples and uses this to categorize the new examples. SVM's are efficient enough to perform even the non-linear classification with the help of kernels. These implicitly map the inputs into high dimensional feature spaces.

MLP is a feed forward artificial neural network model. It maps the input data onto a set of appropriate output by constructing a network of nodes. These nodes are arranged in multiple layers in a directed graph with each layer fully connected to the next one. Except the layer with input nodes all the other nodes are neurons or processing elements with a non-linear activation function. Back propagation, a supervised learning technique, is used by MLP for training the network. MLP is an extension of the standard linear perceptron that can distinguish linearly non-separable data.

Bayesian classifiers assign the most likely class to a given example described by the respective feature vector. Learning such classifiers greatly simplifies the assumption that features are independent of the given class, that is, $P(X|C) = \Pi P(Xi \mid C)$, where $X = (X_1, X_2, \dots X_n)$ is a feature vector and C is a class.

The Naïve Bayes classifier is successfully used in practice often in competition with more sophisticated techniques despite its unrealistic assumption. Naïve Bayes has been effectively used in applications like text classification, medical diagnosis, performance management etc. In chapter 3 machine learning techniques have been applied to cervical cancer demographic data. The same techniques hold in this chapter. The description about the algorithms is not concentrated more whereas the data is explained and the results are projected for a better understanding of the applicability of the algorithms.

4.5 EXPERIMENTS AND RESULTS

In chapter 4 the research is mainly concentrated on the factors that lead to the stage of cervical cancer. Comparative study made on few traditional algorithms with respective to the staging data. The dataset of 203 cervical cancer patient cases is considered in this chapter. This dataset 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: lesion (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. 4 stages have been defined. Each stage again has sub divisions. In this research only the major 4 stages have been considered.

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.

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. ROC curve plot is compared for the performance evaluation of the algorithms along with the RMSE values. Accuracy of a classifier is measured by the area under the ROC curve. ROC is a plot of the true positive rate against the false positive rate for different cut points. Based on this measure the area is computed to analyze the accuracy of the classifier.

Accuracies of different classifiers over the cervical cancer staging data has been analyzed and compared as shown in table 4.1

Table 4.1 Comparative study of classifiers

	AUC					
	J48	**SVM**	**FuzzyRoughNN**	**NN**	**NaiveBayes**	**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*

As can be perceived from the results projected in table 4.1, 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 with a projected accuracy. ROC ranges between 0.86 and 0.95 for the stages of the cancer in J48 with less RSME as low as 0.2284. The efficacies of the algorithms are projected in table 4.2.

Table 4.2: Comparative analysis with/without cross validation

	Accuracy				
Methods Used	**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

Without cross validation the accuracies are high which yield an accurate result as shown in table 4.2. FuzzyRoughNN does not give the necessary result hence can be discarded from the study as it not applicable to any of the datasets. J48 and MLP have projected similar results.

The limitation of MLP is it acts as a black-box. In the further chapters work has been done in the direction of breaking the black box nature from MLP. J48 generates decision tree with the help of information gain concept.

Table 4.3 Sensitivity measures

	Sensitivity			
Methods used	**Stage I**	**Stage II**	**Stage III**	**Stage IV**
J48	0.857	0.710	0.926	0.8
SVM	0.143	0.789	0.941	0.2
FuzzyRoughNN	1	0.421	0.638	0.066
NN	0.5	0.632	0.971	0
NaiveBayes	0.571	0.737	0.956	0.467
MLP	0.571	0.711	0.926	0.467

Table 4.3 shows the sensitivity measures for cervical cancer staging data over a few traditional machine learning algorithms. Sensitivity analysis has several different purposes. The main purpose is in analyzing the performance of the model being applied. This measure helps in identifying the faulty percentage analyzed.

Table 4.4 Specificity measures

	Specificity			
Methods Used	**Stage I**	**Stage II**	**Stage III**	**Stage IV**
J48	0.968	0.960	0.922	0.963
SVM	0.983	0.9	0.776	0.973
FuzzyRoughNN	0.893	0.775	0.857	0.990
NN	0.971	0.924	0.671	0.994
Naivebayes	0.973	0.940	0.835	0.978
MLP	0.968	0.935	0.849	0.962

Table 4.4 gives the specificity values for the same. The results project the performance of the algorithms and one can easily analyze the performance of the different algorithms in the respective stages of the cancer. Specificity and sensitivity measures are used to analyze the performance of the models executed. The measures indicate the how best the model is able to fit the data. It indicates the percentage that could be analyzed truly. The results give a clear indication of the performance of the models.

The dataset has been analyzed with other studies on the similar ground. The results of the same are shown in table 4.5. The results 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.

Table 4.5: Comparative Analysis in similar studies

	Methods Used	**Sensitivity**	**Specificity**	**Accuracy**
Pabitra, 2001(Staging data, clinical)	RoughSet ID3 C4.5	--	--	*81.03* *82.74* *80.2*
Jena, 2005 (MR images of cervical cancer)		*59.26*	*61.54*	*60.95*
Seung Hee Ho, 2004 (demographic, environmental and genetic factors)	CHAID Logistic Regression	*64.00* *40.80*	*77.83* *88.70*	*72.96* *71.83*
Bustanur Rosidi, 2011 (Cervix cells)	Labeled Color Intensity Distribution	*78.6*	*75.9*	*77.0*
Proposed Method, 2014 (Staging Data, Clinical)	C4.5 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*

Accuracy of C4.5 is better as compared to the accuracy achieved so far in similar studies. 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 4.1. The pruned tree generated by J48 is shown in table 4.5

Table 4.5 gives a bird eye on the different algorithms that are applicable and used in analyzing the risk of cervical cancer and spread of the cancer using imaging and textual analysis. This would be helpful to the research community to identify the performance and usability of the algorithms. Inspite of all this survey it is still the nature of data that plays prominent role in the analysis and applicability of the algorithms.

Table 4.5 presents the pruned tree generated by J48. It also gives the instance based classification of the data.

Table 4.6 Pruned Tree by J48

```
Para(u) = N

|  Para(f) = N: Moderate (29.0/7.0)

|  Para(f) = Y

|  |  Va(u) = N: Low (18.0/6.0)

|  |  Va(u) = Y: Moderate (5.0)

Para(u) = Y

|  BB(s) = Y: High (132.0/6.0)

|  BB(s) = N: Risky (19.0/7.0)
```

The decision tree generated by J48 algorithm is shown in figure 4.2

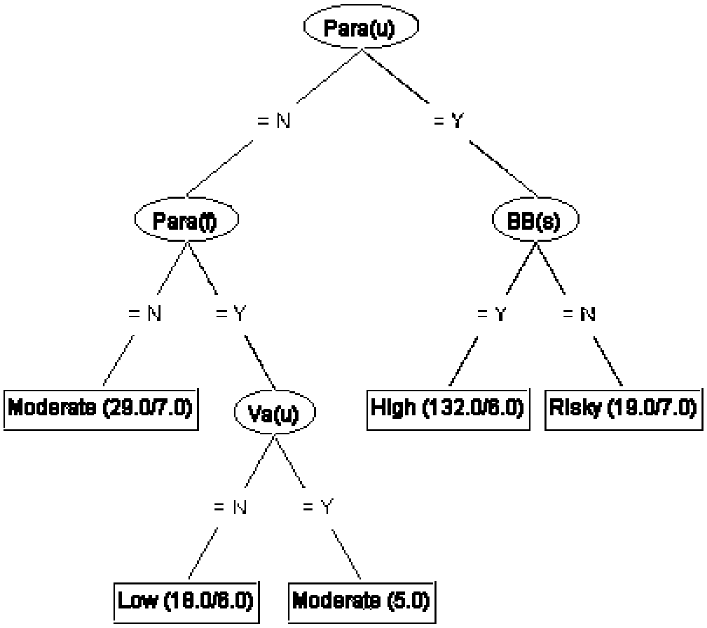

Figure 4.2: Tree for staging of cancer data

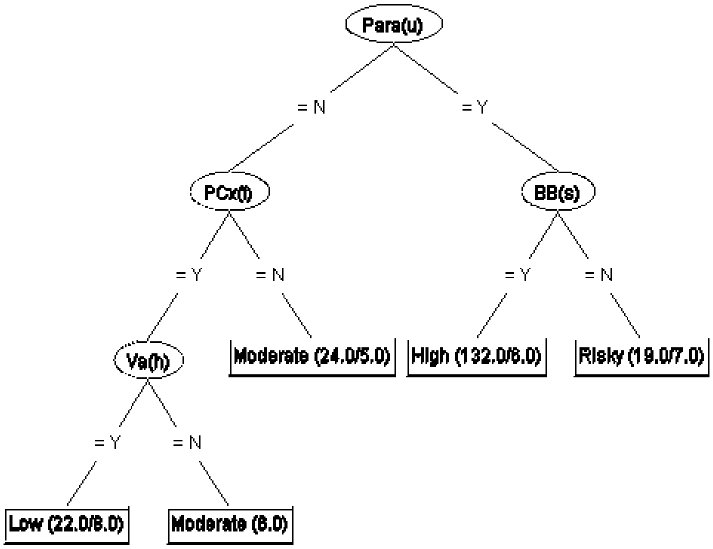

Figure 4.3: Another view of the decision tree on staging of cancer data

Both figures 4.2 and 4.3 give the same decision, Para(u) is the splitting attribute for the construction of the decision tree. If vagina is health then stage is last stage, if the virus is spread to the upper part of the vagina then the stage is moderate ie., 3rd stage. Both the views the tree gives a correct result.

The rules generated by J48 algorithm for the different stages of the cancer 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

Para(u) is the most influential attribute in identifying any stage of the cervical cancer. This information would be helpful to the medical practitioners in identifying the stage of the cancer very early in order to advise a proper treatment or test or medication to the patient.

4.6 CONCLUSION

The chapter introduced the concept of identifying the stage of cancer. Cervical cancer spreads very rapidly if not diagnosed at an early stage. The introduction talks about the issues related to the spread of the cancer, ways in which one can diagnose, cost that is incurred in the diagnosis. In order to reduce the risk of the spread of the cancer and also in order to provide a cost effective solution this study has been made. With the aid of the machine learning algorithms a medical practitioner can identify the stage of the cancer with a few symptoms. If the patients have any lesions found in the vagina then they need to immediately visit the doctor. This has been brought out as a result of the study.

The comparative study of multiple classifiers identifying the stage of cervical cancer using a dataset of size 203 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.

In conclusion it can be perceived 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.