

# Harnessing the Power of Decision Tree approach in Machine Learning for Cervical Cancer Stage Prediction using See5 and SIPINA Sunny Sharma

Department of Computer Science, Guru Nanak Dev University, Amritsar, Punjab, India

## ABSTRACT

Around the globe Cervical cancer is the biggest reason of cancer deaths in women. It influences the cervix in the female regenerative system which prompts death. The decision tree machine learning approach helps to identify the stages of cervical cancer. Decision tree categorize the stages of the cervical cancer in hierarchical decision making system approach which guide the oncologist to take decision on stages of cervical cancer, which safes human life. The proposed methodology use the research data set obtained from http://www.igcs.org & leads the prediction towards the stages of cervical cancer using See5 and SIPINA tool.

Keywords: Cervical Cancer prediction, Machine Learning, See5, SIPINA, Decision Tree, C5.

#### I. INTRODUCTION

The body is made up of lots of living cells. Normal body cells grow, partition into new cells, and pass away in an orderly manner. Cancer begins when cells in a part of the body start to grow out of control. Cancer cell growth is diverse from normal cell growth. Instead of dying or pass away, cancer cells continue to grow and form new, unusual cells. Cells become cancer cells because of damage to DNA structure. We can say cancer is one of the syndrome in which the cells are partitioned & replicated in uncontrolled way. Cervical cancer is one of the most affecting cancers in women worldwide now these days. Its rate of occurrence is around 80% in low & middle income countries or in low socio-economic groups of countries & around 20% in higher income countries or in developed nations. The main problem with cervical cancer is that it cannot be detected as it doesn't show any symptoms until the final stages normally [5] [7].

The machine learning is the technique in which decision boundaries are explored. The Rule base mining technique and Decision trees plays vital role for decision making as well as helpful in machine learning. Another major advantage of using decision tree approach is the white box nature of this approach. It becomes easier to analyse and comprehend the course of decision making process.

#### **II. METHODS AND MATERIAL**

#### A. Decision Trees

Decision Tree depicts the hierarchical decision approach of the problem; with root node & other leaf or internal nodes. It has outgoing edges rather than incoming edges. Internal nodes are called testing nodes & leaf nodes are called terminal nodes as well as decision nodes here in experimental data sets the leaf nodes are the stages of cervical cancer. The ability of Decision tree is that it can illustrate the decision among different attributes. [9][11].

#### **B.** Introduction to Sipina

There are few data mining approaches or tools which have crucial problem to handle large databases but power of SIPINA is that it can deal with large databases easily and ascertain the hidden information in large databases. SIPINA has data mining capability as well as machine learning capabilities. it is free for all kind of activities. SIPINA especially intended to decision trees induction or to do classification, using supervised learning. SIPINA is incorporated with dedicated classification trees algorithms like GID3, ID3, ASSISTANT 86, C4.5, Improved CHAID, One Vs All Decision Tree etc. it also has some other mining capability through Rule Induction, Neural Network, Discriminant Analysis, Decision List etc. we can use any one of them. Experiment is done on cervical cancer data bank. Small sample of data set of 237 contains seven stages and 12 attributes. [17]

#### C. C5 Algorithm Implementation Of SEE5

Quinlan's C5.0 algorithm is widely used for classification process. Algorithm primarily focuses on constructing a decision with the identification of most important attributes from the supplied/identified data-set. Once the attribute is finalized from current node, corresponding child nodes are then generated. There after best attribute of a node can be selected. There are few options present over See5 tool like, Boosting is to generate several classifiers (decision trees or rule-sets) instead of one. On classifying a new case, each classifier supports its predicted class and then the support is evaluated to determine the final class. In the first step, a single decision tree or rule-set is constructed as before from the training data. This classifier will usually make mistakes on some cases, like here the first decision tree, gives the wrong class for 14 cases in sequence data. Other classifier is constructed giving more consideration to the cases. Thus the classifier will provides results variation from the earlier classifier. Errors induced are again rectified by another classifier. It continues for defined iterations/trials and halts once extremely correct classifies is achieved [18], [19].

Winnowing is a mechanism to separate the useful attributes from useless attributes. It provides option to select among the predictors and have an edge to create a suitable decision-tree. However, it's time intensive task and primarily suitable for bigger application domain. [18], [20].

In Advanced pruning technique a massive tree is first allowed to grow to fit the data closely after that it's pruned i.e. error causing segments are removed. Every sub-tree undergoes pruning then replacement by a leaf or sub branch is decided and then a global stage evaluates performance of the tree as a unit. [18], [20].

### D. Stages Of Cervical Cancer

### 1) Stage 0-Carcinoma in Situ:

Stage 0 is carcinoma in situ i.e origin of abnormal cells in the inner-most lining of the cervix. They become cancer & affect nearby customary tissue.

#### 2) Stage I

At this stage presence of cancer is in cervix only.

- Stage I-A: with the help of microscope cancer can be seen in cervix tissues. Further detail is expressed as.
  - Stage I-A (1): cancer depth is not more than 3 millimeters and it is not greater than 7 millimeters wide in tissues.
  - Stage I-A (2): cancer depth is greater than 3 but not greater than 5 millimeters, but not more than 7 millimeters wide.
- Stage I-B detail is expressed as.
  - Stage I-B (1): the depth is greater than 5 millimeters & greater than 7 millimeters wide.

Stage I-B (2): the not more then 4 centimeter wider cancer can be seen without the help of microscope.

# 3) Stage II

At this stage cancer has spread beyond the cervix but not towards the pelvic wall or towards the inferior third of the vagina. With deep penetration of cancer further detail can be expressed as.

- Stage II-A: Cancer increased from cervix towards the superior two third of the vagina, but not towards tissues of the uterus.
  - Stage II-A (1): without the help of microscope tumor of not more than 4 centimeters can be seen.
  - Stage II-A (2): without the help of microscope tumor of more than 4 centimeters can be seen.
- Stage IIB: Cancer stretched from cervix towards the tissues of the uterus.

### 4) Stage III

At this stage cancer stretched towards the inferior third of the vagina, Pelvic wall can cause kidney troubles. With deep penetration of cancer further detail can be expressed as.

- Stage III-A: Cancer has stretched towards the inferior third of the vagina but not towards pelvic wall.
- Stage III-B: Cancer has stretched towards the pelvic wall, tumor has become huge enough to chunk the ureters which increase the size of kidneys or can stop kidneys working.

#### 5) Stage IV

At this stage cancer stretched towards the bladder/rectum, or other parts of the body.

- Stage IV-A: Cancer stretched towards nearly connected organs like rectum/bladder.
- Stage IV-B: Cancer stretched towards other parts of the body like liver/lungs/bones/distant lymph nodes [3]

Common types of treatments for cervical cancer are as follow: Surgery/Radiation therapy/ Chemotherapy & combination of them

#### E. Literature Survey

In 2009 A. Satija et al. revealed the statistics of cervical cancer in India. It is primarily caused by human papillomavirus (HPV) infection with the vaccination much progress has been made in the prevention and control of cervical cancer [2]. In 2009 G. Jayalalitha et al. discussed technique of grading cervical cancer images according to the cell formation of tissues. They expressed the use of Box Counting Method (DB) and Harfa Programme software to detect the fractional dimension and calculate the variation of intensity and texture complexity of cancer cell images [4]. In 2009 the C. Todd et al. proposed computer assisted algorithm for the classification of cervical cancer using digitized histology images of biopsies. Texture analysis of the nuclei structure is very important for the classification of cervical cancer histology[6]. In 2010 M. Ross et al. explore the focus towards the use of histology images for the classification of cervical cancer [8]. In 2010 S.Allwin et al. proposes approach which use textural properties to classify the various malignancies in cervical cyto images [10]. In 2011 C. Balleyguier et al. proposed the guidelines for staging & follow up of patients which suffered from uterine cervical cancer & provides the radiologists with a framework & expressed the importance towards adequate patient preparation,

protocol optimization and MRI reporting expertise are essential to achieve high diagnosis accuracy [11]. In 2012 **M. Singh et al.** use the decision trees to predict the protein classes through see5 tool and express the importance of decision tress in bioinformatics. In 2014 **S. Sharma et al.** describe the importance of evolutionary multiobjective optimization algorithms in bioinformatics [15]. The methodology used for the research experiment is expressed as.



#### **III. RESULTS AND DISCUSSION**

The Pearson correlation between various features like NodePET, ClinDiameter, MRIVol, UterineBody, Status, RelPrimary, Relpelvic, RelAbdo, RelSupraclav, RelDistant is obtained and shown in the table. Which describe that RelAbdo, RelDistant, Status, Histoloogy & RelPrimary plays vital role in decision making. Correlation graph for various features is shown in Figure 1.



Figure 1. Correlation between various features

	Goodness of split	Correlation	Accept or Reject
MRIVol	0.13688986	0.0625	
ClinDiameter	0.11381233	0.0599	
RelSupraclav	0.10874788	0.0193	
UterineBody	0.06534168	0.0288	
Histology	0.05033587	0.0094	
RelDistant	0.00000000	0.0000	
RelAbdo	0.00000000	0.0000	
RelPelvic	0.00000000	0.0000	
RelPrimary	0.00000000	0.0000	
NodePET	0.00000000	0.0000	
Status	0.00000000	0.0000	

The Goodness of split and correlation values of each attribute for decision making is shown in Figure 2

Figure 2: Goodness of Split with Correlation of Attributes

The suggested attribute for decision tree by SIPINA is MRIVol and the suggested value for the decision node is 28.26. The number of stages predicted on the basis of this value is shown in Figure 3.

Split suggestion for "MRIVol"				
	< 28.26	>=28.26		
Stage-1a	1	0		
Stage-1b	35	34		
Stage-2a	20	14		
Stage-2b	24	56		
Stage-3a	4	5		
Stage-3b	2	41		
Stage-4a	0	1		

Figure 3: Suggested Root Node in SIPINA with Decision Value

The total 237 data set cases discription on the basis of Cervical Stages as well as on the basis of Attributes with individual percentages is expressed in Figure 4.

	st	ages (-)		
Values	Strength	Local Dist.	Global Dist.	Recal
Stage-1a	0.00	1 (0%)		-
Stage-1b	0.00	69 (29%)	-	-
Stage-2a	0.00	34 (14%)		
Stage-2b	0.00	80 (34%)		
Stage-3a	0.00	9 (4%)	-	
Stage-3b	0.00	43 (18%)		
Stage-4a	0.00	1 (0%)		
	His	tology ( - )		
Values	Strength	Local Dist.	Global Dist.	Recal
SCC	0.00	204 (86%)	-	
Endometroid	0.00	26 (11%)		1
Clearcell	0.00	5 (2%)	-	2
Serous	0.00	2 (1%)	2	2

Figure 4 : Cervical Stages and Attribute contribution

The training Vs. Testing Lift curve is shown in Figure 5 which describe the Training, Perfect Training and testing graph.



Figure 5: Lift Curve Training vs. Testing

Similarly the scoring target curve of training and testing is shown in Figure 6.



Figure 6: Scoring Matrix

Based on this scoring as well as on confusion matrix the accuracy on SIPINA is determined. By taking the random Sample of 119 data sets the accuracy on remaining data sets is calculated as 26.89%. Similarly by taking random sample of 117 data sets accuracy on training data sets is calculated as 59.82%. Again by considering whole datasets the accuracies are calculated 69.9%. Similarly for the data set of 237 patients with 12 features, the accuracy of the different techniques was calculated on See5 as shown in Figure 6. The C5 algorithm gives 67.5% accuracy using advance pruning option.



Figure 7. Shows the Maximum accuracies obtained using See5

#### International Journal of Scientific Research in Science, Engineering and Technology (ijsrset.com)

Decision tree obtained from SIPINA is shown in Figure 8. Similarly Decision tree based on various features present in See5 is obtained and it is shown in Figure 9.

## **IV. CONCLUSION**

For treatment of patients or in diagnosis process huge amount of information is processed specially for cancer patient. Oncologist can take the help of decision tree based computerized approach to diagnose their patient. The proposed methodology classifies the stages of cervical cancer using See5 and SIPINA.

### **V. REFERENCES**

- [1] www.igcs.org/professional/Education/treatmentRe sources/CervicalCaDB.html.
- [2] A. Satija, "Cervical cancer in India", South Asia Centre for Chronic Disease, 2009.
- [3] http://www.cancer.gov/cancertopics/pdq/treatment /cervical/Patient/page2.
- [4] G. Jayalalitha and R. Uthayukumar, "Recognition of Cervical cancer based on Fractal Dimension", International Conference on Advances in Recent Technologies in Communication and Computing, 2009.
- [5] K.Jayant, R.S Rao, "Improved stage at diagnosis of Cervical cancer with increased cancer awareness in rural Indian population", International Journal Cancer, 1995, Vol.63, pp 161-163
- [6] C.Todd , Rahmdwati, G.Naghdy, "Cervical Cancer Classification Using Gabor Filters", First IEEE International Conference on Healthcare and Informatics, Imaging and Systems Biology, 2011
- [7] Cevical Cancer Overview", American Cancer Society, http://www.cancer.org/acs/groups/cid/documents/ webcontent/003042-pdf.pdf.
- [8] Montse Ross and Rahmadwati, "Classification Cervical Cancer Using Histology Images", Second international Conference on Computer Engineering and Application, 2010.
- [9] B. Bergeron, "Bioinformatics Computing", pp 257-270, 2002.
- [10] S.Allwin and S.Pradeep Kumar, "Classification of stages of Maligancies using Textron signature of

Cervical Cyto Image", Computational Intelligence and Computing Research (ICCIC), 2010.

- [11] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA pp 279-322, 2003.
- [12] C. Balleyguier et al., "Staging of uterine cervical cancer with MRI", Journal on European Radiology, 2011
- [13] I. Friedberg, "Automated Protein Function Prediction- the Genomic Challenge", Briefings in Bioinformatics, vol 7, no.3, pp 225-242
- [14] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames C. Kesmir, H. Nielsen, H.H. Stærfeldt, K. Rapacki, C. Workman C.A.F. Andersen, S. Knudsen, A. Krogh, A.Valencia and S. Brunak, "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features", Journal of Molecular Biology, vol. 319, issue 5,pp 1257-1265, 2002.
- [15] MS Sharma," A Review towards Evolutionary Multiobjective optimization Algorithms", An International Journal of Engineering Sciences, Vol13/37-Vol13, Vol. 3, Issue December 2014.
- [16] M. Singh, G. Singh, S. Sharma," Human Protein Function Prediction from Sequence Derived Features using See5 ", International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012
- [17] http://eric.univ-lyon2.fr/~ricco/sipina.html
- [18] H. Wei-Feng, G. Na, Y. Yan, L. Ji-Yang, Y. Ji-Hong, "Decision Trees Com-bined with Feature Selection for the Rational Synthesis of Aluminophos-phate AlPO4-5", *National Natural Science Foundation of China*, vol 27, no.9, pp 2111-2117, 2011.
- [19] I. Friedberg, "Automated Protein Function Prediction- the Genomic Chal-lenge", *Briefings in Bioinformatics*, vol 7, no.3, pp 225-242
- [20] http://rulequest.com/see5-info.html



Figure 8: Decision Tree obtained from SIPINA

1181



Figure 9: Decision Tree obtained from See5