Predicting Cervical Carcinoma Stages Identification using SVM Classifier

Chandra J [#1], Nachamai.M [*2], Anitha S Pillai [#3]

[12]*Associate Professor, Department of Computer Science, Christ University, Bangalore, Karnataka, India.*

[3] *Professor & Head, Department of MCA, Hindustan University Chennai, Tamil Nadu, India*

***Abstract*— Cervical Cancer is one of the most specific cancers among women global. This paper is a candid attempt to cover a small gap in cervical cancer research practice in India. Longitudinal studies are one area under which more resourceful data is required to treat on the patients affected. This work would prove evidential for an effective intervention and systematic evaluation to assess the impact short term and long term carcinoma. The International federation of Gynaecology and Obstetrics staging of cervical carcinomas are into four stages: Invasive cancer, Clinical lesions, Pelvis, and True pelvis. In Machine learning, research has concrete proof that support vector machine is a better approach to learn any data base, SVM classification method performance is superior to any other commonalities like bagging, boosting. The upshots corroborated to be tangibly supporting the original diagnosis given by the IGCS. This modus espoused has proven that it is indubitable investigative support for diagnosing of cervical carcinoma staging. The results of staging classification can be used by the clinical specialists to decide upon the remedial course of treatment of carcinoma.**

Keywords--

Cervical Cancer (CC), Support Vector Machine (SVM), federation of Gynaecology and Obstetrics staging (FGOS), low and middle income countries (LMICs),

I. INTRODUCTION

Cervical cancer is wide spread, making it complex to identify and detect among huge residents. The classification of cervical cancer has been a challenging task for research. CC is humanity exemplifies health disparity, as their charges are superior in LMICs and in low socio-economic groups within country. Around 80% of global cervical cancer cases are in LMICs. Cervical cancer causes loss of active life both due to early death as well as prolonged disability. Among women aged 25-64 years, who tend, in India, to be the sole caretakers of the house and family, and in some cases important contributors to the family earnings, this humanity weight poses a deep financial burden on families as well the country (as in National Commission on Macroeconomics of Health, 2005) and moreover, the medicinal costs that are incurred by families due to cervical cancer. Most cases in growing countries are diagnosed at complex stages when treatment is lavish but forecast poor.

Machine learning is the process of analysing data from different perspectives and summarizing it into useful information. Data are any facts, text or numbers that can be processed out. The information is useful to increase the revenue and reduce the costs of the organization. Machine learning tools allow the users to analyse the data from different dimensions. A data mining system contains data, information and knowledge to extract these data and information. CC remains as a leading cause of morbidity and humanity for women worldwide. Current predictable treatment includes radiotherapy, however a substantial percentage of patients have insensitive tumors presenting mechanism that allow them to escape from the special effects of energy. There is a need for healthy biomarkers to optimize treatment and to use more violent restorative agents, mainly in budding countries where a large number of women's are previously CC patients and sympathy is closely linked to late judgment of neoplasias and failure of healing. Viral oncoproteins have the ability to modify the appearance rate of specific miRNAs that could be associated to radio-response in resistant cells. This review has emphasized the role of miRNAs that could regulate radio-resistance mechanisms; hence, the evaluation of those miRNAs as potential biomarkers opens up new horizons for CC prediction and treatment.

CC is a difficult disease in which progression from normal infected cells to pre-neoplasic lesions, metastatic disease, and clinical response to radiotherapy, are affected by several factors, not only associated to HR-HPV infection and E6/E7 oncoprotein interaction with cellular components. In this scenario, miRNA could be a key fact in the explanation of these complex phenotypes, such as radio resistance. Hopefully, in the near future we can bring more discoveries on miRNAs science to develop molecular markers associated to cancer progression, clinical outcome, or probably to use them not only to block E6 and E7 expression in cervical cancer cells, but also to reprogram the cancer cell.

II. MATERIALS AND METHODS

There are a number of Data mining techniques like classification, clustering, SVM, multiple predictive models etc. Here, there are few techniques of data mining, which are considered as important for analysing the bank data. SVM was developed by Vapnik [1] and his colleagues in AT & T Bell laboratories. SVM consists of a set of related supervised

learning methods. SVM is a valuable method for data categorization. The SVM classification involves with training and test data sets. Each instance in the training set contains one "target value" and several "attributes". SVM Ensemble classifier is a collection of several SVM classifiers whose individual decisions are combined to classify the test samples [2].An ensemble shows better performance than individual classifiers from which it is constructed [3]. The SVM Ensemble classification prediction includes two levels: classifier construction and the usage of the classifier.

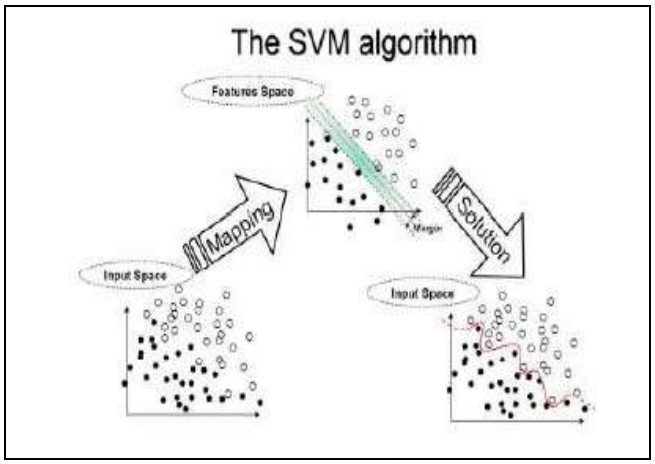

Fig. 1 SVM Classifier

The fig 1 shows the SVM classifier and it is constructed from the training set. Each example in the training set is unspecified to belong to a predefined class, as gritty by the class attribute label. Using WEKA knowledge flow and classifier selection, a boosted SVM Ensemble is created. The created model is used for further prediction. The later involves the use of SVM Ensemble built to predict or classify the output. The processes start by training an SVM classifier with a less imbalanced subset of data, and then classify the entire training data set with the SVM to identify the incorrectly classified examples. The trained models are aggregated using majority voting to obtain a collective outcome. The main objective of the SVM is to predict target value of data instances in testing set with better accuracy. It uses the information of entire dataset. SVM can perform well on dataset that are having many instances. There are a number of Data mining techniques like clustering, classification, SVM, prediction etc. Here, there are few techniques of data mining, which are considered as important for analysing the bank data. Support Vector machine was developed by Vapnik [1] and his colleagues in AT & T Bell laboratories. SVM consists of a set of related supervised learning methods. SVM classification is an example of supervised learning. There is no upper limit for attribute count. SVM is a useful technique for data classification. In SVMs, a data point is viewed as a p-dimensional vector. If these p points are separable using (p-1) dimensional hyperplane are called linear classifier. Let **Z** be the dataset for linear SVM. It contains a set of points of the form

$$Z = \{ (X_i, Y_i) | X_i \in R^p, Y_i \in \{-1,1\} \}_{i=1}^{n} \quad \text{----------(1)}$$

Where Y_i is either 1 or –1 and it indicates the class to which the point $\mathbf{X}_i$ belongs.

SVM ensemble is a collection of several SVM classifiers. The decision to classify the data is obtained by combining the decision functions of all the individual classifiers. The classification result is predicted through the aggregation of many individual SVM classifiers. The classifiers should be different from each other. Same set of classifiers will produce the same result. But for more accurate results the classifiers are needed to be different in some way from other. The final result of an ensemble is obtained by combining the classifiers. An ensemble of several individual SVM classifiers is expected to achieve better performance than single SVM. With the help of SVM classifier, there are several research happens on decision-making, prediction making, and knowledge discovery etc. It is successfully applied to real-world applications like face recognition, pattern recognition, text classification, Spam categorization, financial research etc. In [2], the SVMs were used for predicting bankruptcy and accuracy was generated by SVM, through this experimental work, the author found that the SVM predictor is better than other methods. It is considered as a good classifier and predictor tool. Many researches are doing research on financial sector using SVM. In [3] Harris Druker et al., has implemented SVMs for e-mail classification. The author used both SVM and classification algorithms to classify e-mail to spam and non-spam. In [4], Guojun Zhang has developed a high-performance classifier using ensemble classifier algorithm based on SVM. SVM is used for doing many research with bankruptcy prediction, fault diagnosis, municipal revenue prediction, handwriting recognition, financial forecasting were done with the help of Support Vector Machine.

III. RESULT

The SVM classifier was implemented using Weka 3.7.4 an open source data mining tool. In Weka different learning methods can be applied to the dataset to extract useful information about the data. The Lib-SVM was used for execution. The methods classified the four stages with stage I expending the criteria stroma length more than 3 mm in depth and no wider than 7 mm diameter, Stage II captivating the clinical lesions greater than 4 cm, Stage III presumed pelvic sidewall dimensions and stage IV true pelvis measuring abnormality of the bladder size.

A. Data set Description

The data set used for this work is International Gynaecologic Cancer Society sponsored open dataset with 237 patient data which is the Pre-treatment clinical as given by (FIGO) staging, Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) performed. The data is available online as an excel file, with 237 rows and 8

parameter values for the classification. This modus espoused has proven that it is indubitable investigative support for diagnosing of cervical carcinoma staging. The results of staging classification can be used by the clinical specialists to decide upon the remedial course of treatment of carcinoma.

The FIGO staging of cervical carcinomas are into four stages: Invasive cancer, Clinical lesions, Pelvis, and True pelvis. The SVM classifier was implemented using weka-2.7 which classified the four stages:

Stage I: Expending the criteria stroma length more than 3 mm in depth and no wider than 7 mm diameter.

Stage II: Appealing the clinical lesions greater than 4 cm.

Stage III: Presumed pelvic sidewall dimensions.

Stage IV: True pelvis measuring irregularity of the bladder size.

B. *Performance Measures*

Performance metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future. For evaluating the performance of a classifier confusion matrix, accuracy values etc are used. The accuracy is defined as,

Accuracy = TP+TN / (TP+FP+TN+FN) ----------(2)

Where TP, FP, TN and FN are the numbers of true positive predictions, false positive predictions, true negative predictions and false negative predictions, respectively. The Precision, Recall, F-Measures etc are also listed from the WEKA classifier performance evaluator.

Scheme: ***weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7"***

Relation: ***CervicalCancer_Narayan_IGCS2***

Instances: ***237 Attributes:*** ***6***

Histology

FIGO

nodePET +

Clin.diameter cm

MRIVol cc

Uterine Body

Test mode: Evaluate on training data

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM) ***Fig 2. Experimental result on training data set***

Correctly Classified Instances 234	***98.7342 %***	
Incorrectly Classified Instances	***3***	***1.2658 %***
Kappa statistic		***0.9687***
Mean absolute error		***0.0127***
Root mean squared error		***0.1125***
Relative absolute error		***3.0877 %***
Root relative squared error		***24.8733 %***
Coverage of cases (0.95 level)		***98.7342 %***
Mean rel. region size (0.95 level)		***50 %***
Total Number of Instances		***237***

Time taken to build model: 0.03 seconds

=== Detailed Accuracy By Class ===

TP Rate	***FP Rate***	***Precision***	***Recall***	***F-Measure***	***ROC Area***
1	***0.044***	***0.983***	***1***	***0.991***	***0.978***
0.956	***0***	***1***	***0.956***	***0.977***	***0.978***
0.987	***0.031***	***0.988***	***0.987***	***0.987***	***0.978***

Fig. 3 Experimental result on accuracy

The experiment was carried out for different policies in the same way and found that the SVM ensemble performs better than the single classifier. Fast and accurate classifier is used for making investment prediction. In fig 2 and fig 3 shows the experimental result on cervical cancer data set using SVM classification. From the experimental result, it is found that SVM can be used as a predictor for identifying different stages involved in CC. Based on the stages treatment can be given.

IV. CONCLUSIONS

To summarize the developed method, CC is a worldwide public health problem among women. To improve the control of cervical cancer new diagnostic and therapeutic strategies are required. Integration of immunology and proteomic biotechnology with computational tools like SVM has accelerated the understanding of the genetic and cellular basis of many cancer types. Finally the SVM classifier is used for classification. Cervical cancer is a worldwide public health problem among women. To improve the control of cervical cancer new diagnostic and therapeutic strategies are required.

Integration of immunology and proteomic biotechnology with computational tools like SVM has accelerated the understanding of the genetic and cellular basis of many cancer types CC can be cured and treated when found.

REFERENCES

[1] Bin Linghu; BingYu Sun; , "Constructing effective SVM ensembles for image classification," Knowledge Acquisition and Modeling (KAM), 2010 3rd International Symposium on , vol., no., pp.80-83, 20-21 Oct. 2010.

[2] Fumera, G.; Roli, F.; , "A theoretical and experimental analysis of linear combiners for multiple classifier systems," Pattern Analysis and Machine Intelligence, IEEE Transactions on , volume 27, Issue No.6, pp.942-956, June 2005.

[3] Bhardwaj, M.; Gupta, T.; Grover, T.; Bhatnagar, V., "An efficient classifier ensemble using SVM," Methods and Models in Computer Science, ICM2CS 2009. Proceeding of International Conference on , pp:240-246, 14-15 Dec. 2009.

[4] Cheolkon Jung; Jiao, L.C.; Yanbo Shen; , "Ensemble Ranking SVM for learning to rank," Machine Learning for Signal Processing (MLSP), pp.1-6,PP:18-21,.2011

[5] Kesheng Lu; Lingzhi Wang; , "A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction," Computational Sciences and Optimization (CSO), Fourth International Joint Conference on ,pp.1343-1346,15-19,April2011.

[6] Ming-Hsuan Yang; Moghaddam, B, "Support vector machines for visual gender classification," Pattern Recognition, 2000. Proceedings. 15th International Conference on, pp.1115-1118,2000.

[7] Ma Chao; Chen Xihong; , "A New Algorithm of Support Vector Machine Ensemble and Its Application," Intelligent Human-Machine Systems and Cybernetics (IHMSC),2nd International Conference on , pp.225-229, 26-28 Aug. 2010.

*[8] Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research **11**: 169–198. doi:10.1613/jair.614.*

*[9] Polikar, R. (2006). "Ensemble based systems in decision making". IEEE Circuits and Systems Magazine **6** (3): 21–45. doi:10.1109/MCAS.2006.1688199.*

*[10] Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review **33** (1-2): 1–39. doi:10.1007/s10462-009-9124-7.*

[11] Kuncheva, L. and Whitaker, C., Measures of diversity in classifier ensembles, Machine Learning, 51, pp. 181-207, 2003

[12] Chawla, Nitesh V (2004). Learning Ensembles from Bites : A Scalable and Accurate Approach. Journal of Machine Learning Research, pp. 421–451.

[13] C.Cortes,V.Vapnik ,Support vector network, Machine Learning, pp.273–297,1995.

[14] Fumera, G, Roli, F ,"A theoretical and experimental analysis of linear combiners for multiple classifier systems" Pattern Analysis and Machine Intelligence ,IEEE Transactions , vol.27, no.6, pp.942-956.

[15]Harris Drucker 'et al," Support vector machines for spam categorization" IEEE Transactions on Neural Networks, 1999.

[16]Kuncheva LI.,Combining Pattern Classifiers. Methods and Algorithms,2004.

[17] Leon Bottou and Chih-Jen Lin, Support vectorMachine Solvers in Large Scale Kernel Machines. Weston editors, MIT Press, Cambridge, MA, pp. 1–28.

[18] Ma Chao, Chen Xihong ,"A New Algorithm of Support Vector Machine Ensemble and Its Application",Intelligent Human- Machine Systems and Cybernetics (IHMSC), 2nd International Conference, 225-229,2010.