ISSN 2321 3361 © 2018 IJESC

Research Article Volume 8 Issue No.5

Constructing a Predictive Model for Detection of Breast Cancer

Fantaye Ayele
MSc Lecturer
Department of Information System
Woliata Sodo University, Woliata Sodo, Ethiopia

Abstract:
Cancer is a big issue all over the world. It is a disease, which is incurable in many cases and has affected the lives of many and will continue to affect the lives of many more. Breast cancer is the second leading cause of cancer deaths in women now a day and has become the most common cancer among women both in the developed and the developing world. Early detection is the most effective way to decrease breast cancer deaths. But early detection needs an accurate and reliable diagnosis procedure that allows doctors to differentiate benign breast tumors from malignant ones without going for surgical biopsy. Hence, construct a predictive model using data mining techniques to identify hidden knowledge and develop a prototype interface for breast cancer that support health professional in their diagnosis decisions and treatment planning measures. For this study, a six-step hybrid knowledge discovery process model is followed, due to the nature of the problem and attributes in the dataset. The classification technique such as, J48 decision tree, Naive Bayes and PART rule induction used to build the models. Performance of the models is compared using accuracy, TPR, TNR, and the area under the ROC curve. J48 decision tree registers better performance with 94.82 % accuracy.

Keywords: Benign, Breast Cancer, classification, Data mining, Knowledge Discovery in Databases, Malignant, WEKA

I. INTRODUCTION

Breast cancer is the second most often occurring cancer next to cervical cancer among women in Ethiopia. It is predictable that around 10,000 Ethiopian women have breast cancer [1]. One out of eight women will develop breast cancer during her lifetime [2]. The diagnosis result of tissue is classified into three classes: normal which represents mammogram without any tumorous cell, benign which represents mammogram showing a cancer, but not formed by cancerous cells and malign which represents mammogram showing a cancer with tumorous cells [3]. The goal of breast cancer diagnostic prediction is to allocate patients to either a benign group that is non-cancerous or a malignant group that is cancerous [4]. There are over 212,000 cases of breast cancer diagnosed in the USA every; in Canada, Australia and United Kingdom the figure is 20,500, 13, 000 and 41, 000 respectively. In Ethiopia, overall one woman in every nine grows breast cancer at one-time in her life [5]. Data mining has been defined as the non-trivial extraction of earlier unknown, implied and potentially useful information from data. It is the science of mining useful information from huge databases. Data mining is one task in the procedure of knowledge discovery from the database [6].

II. STATEMENT OF THE PROBLEM

Cancer is becoming the leading cause for death in the world; over half of them die because of the late diagnosing of the disease [7]. Tikur Anbessa Specialized Hospital is a hospital which have large amount of breast cancer patients than other hospitals in Ethiopia. It is the only cancer referral hospital in Ethiopia, has limited ability to care for cancer patients. The hospital works only with three adult oncologists, who represent all such specialists in the nation, endures a chronic shortage of chemotherapy drugs, Chemotherapeutic medicines are only two radiotherapy machines exist for more than 90 million people and requiring patients to wait up to six months for treatment of their diagnosis[8]. Identifying the best attribute to diagnoses breast cancer is still challenging. Whereas significant efforts are made to attain early detection and effective treatment, but scientists do not distinguish the strict causes of most breast cancer. They do know some of the risk factors initiating breast cancer, such as ageing, genetic risk factors, family history, menstrual periods, not having children, obesity that increases the probability of developing breast cancer in females [4]. Other problem concerning breast cancer is cost and time involved in screening due to the several tests for diagnosis. Various tests are available for predicting breast cancer, but detecting breast cancer in earlier stage is difficult for the doctors, but earlier detection of breast cancer is possible by using data mining techniques. Hence, predictive model estimates easily and a cost effective way for screening breast cancer and may play a pivotal role in earlier diagnosis process and provide effective preventive strategy.
Hence, the motivation of the computer-aided diagnosis systems is to support medical staffs to attain high efficiency and accuracy. This paper using local data were done by applying data mining technique for breast cancer diagnosis. Thus, this study therefore aims to construct a predictive model by using local dataset for breast cancer diagnosis. Hence, this paper attempts to answer the following questions:

- Which attributes are more important to predict the breast cancer?
- Which classification algorithm can be more suitable for the purpose of identifying/predicting breast cancer states?
- How much is the prototype accepted by the domain experts during evaluation of the prototype?

III. OBJECTIVE OF THE PAPER

A. General Objective

The major objective of this paper is to construct a predictive model using data mining techniques to identify hidden

knowledge and develop a prototype interface that assists physician to access the identified hidden knowledge for determining breast tumor states.

B. Specific Objectives

- To identify the important attributes that help to predict breast tumor states of patients.
- To find out classification algorithm that more suitable to build predictive model for breast cancer.
- To evaluate the performance of the models.
- To develop a prototype.

IV. METHODOLOGY OF THE STUDY

This paper was used a Hybrid data mining model which is a six step knowledge discovery process model. Due to the nature of the problem and attributes in the dataset, classification mining task were selected to build the predictive model.

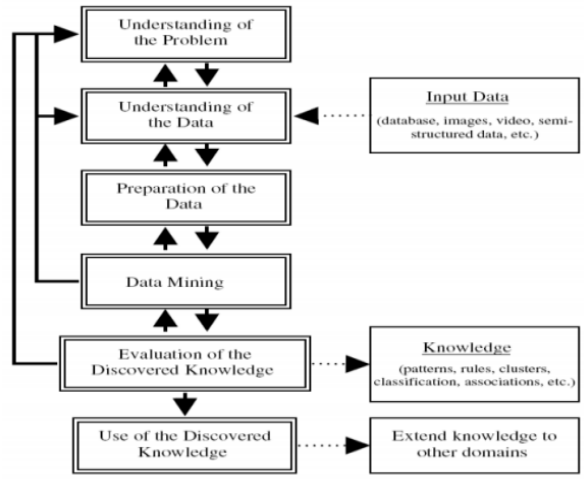


Figure .1. Hybrid-DM Process Model [9]

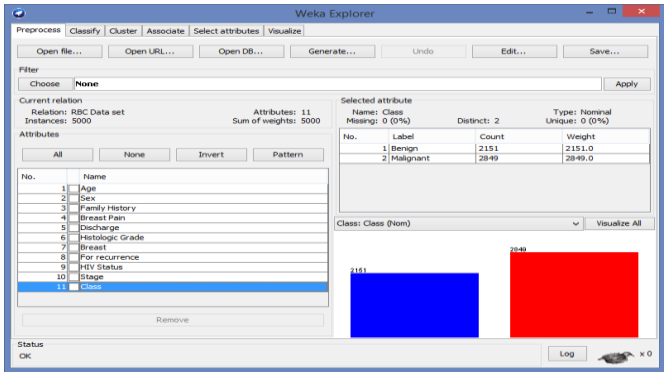
V. EXPERIMENTATION AND ANALYSIS OF RESULT

The purpose of experiments in classification is to find model that is able to predict the breast tumor status of patients as benign or malignant tumor taking selected variables as inputs. This paper incorporated the typical stages that characterize a data mining process.

a) Experimental Design

In this paper, all experiments are done based on the final processed dataset which contains 5,000 instances and 11 attributes. The algorithms used during predictive model building experimentations are found in Weka 3.7.5 version.

Figure 2: Weka Explorer window showing the number of attributes and instances.



b) Model Building

Model building is one of the major tasks which are undertaken under the phase of data mining in Hybrid data mining methodology. To build the predictive model, J48, naive Bayes and PART algorithms are trained and evaluated. For training and testing the classification model the researcher used two methods. The first method is percentage split method, where 75% of the data used as training and the remaining 25% testing. The second method is K-fold cross validation methods the data was divided into 10 folds, some fold is used as testing and the remaining folds are used as training.

Table .1. Type of experiment

Algorisms	Parameters	Experiments	Scenario /methods	Experiments number
J48/ Decision tree	Pruned	75/25 split criteria	1	Exp1
		10 cross fold validation	2	
	Unpruned	75/25 split criteria	1	Exp2
		10 cross fold validation	2	
NB/ Naïve Bayes	Default	75/25 split criteria	1	Exp3
		10 cross fold validation	2	
PART/ Rule induction	Pruned	75/25 split criteria	1	Exp4
		10 cross fold validation	2	
	Unpruned	75/25 split criteria	1	Exp5
		10 ross fold validation	2	

Experimentation I: J48 decision tree with Pruned

Table.2. Summary of Experiments I with J48 pruned decision tree

Exp I (test model)	Accuracy	Time Taken	Tree Size	Leaf Size	W TPR	W FPR	W PR	W RR	W ROC	CCI	ICI
J48 pruned 75/25 percentage split	94.64	0.8	42	27	0.946	0.061	0.947	0.946	0.976	1183	67
J48 pruned 10-fold cross validation	**94.82**	0.11	42	27	**0.948**	**0.058**	**0.948**	**0.948**	**0.98**	4741	259

Key: *Exp: experiment, CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, W: Weighted Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate, FR: F-measure rate.*

Finally, scenario #2 is better than scenario #1 in terms of accuracy, correctly classified instance and ROC curve. Therefore, scenario #2Building pruned decision tree with all attribute of 10- fold cross validation was selected.

Experimentation II: J48 decision tree with UnPruned

Table.3. Summary of Experiments II with J48 unpruned decision tree

Experiment II	Accuracy	Time Taken	Tree Size	Leaf Size	W TPR	W FPR	W PR	W RR	W ROC	CCI	ICI
J48 un pruned 75/25 percentage split	94	0.05	68	46	0.94	0.062	0.94	0.94	0.979	1175	75
J48 un pruned 10-fold cross validation	**94.4**	0.06	68	46	0.944	0.06	0.944	0.944	**0.981**	4720	280

Accordingly, scenario #2Building unpruned decision tree with all attribute of 10- fold cross validation was selected as the best J48 decision tree using experiment 2.

Experimentation III: Naive Bayes with 10-fold cross-validation and 75/30 percentage split

Table .4. Summary of Naive Bayes Experiments

Experiment 3	Accuracy	Time Taken	Av TPR	Av FPR	Av PR	Av RR	Av ROC	CCI	ICI
Naive Bayes 75/25 percentage split	85.6	0	0.856	0.161	0.858	0.856	0.906	1070	180
Naive Bayes 10-fold cross validation	**87.42**	09	**0.874**	**0.147**	**0.877**	**0.874**	**0.918**	4371	629

Based on the evaluation criteria, the classifier correctly classifies patients as Malignant that has actually Malignant with 94% accuracy. Hence, cross fold validation selected as the best Naive Bayes that evaluating the model based on sensitivity and specificity result.

Experimentation IV: PART with Pruned

Table .5. Summary of PART Experiments with pruned

Experiment	Accuracy	Time Taken	W TPR	W FPR	W PR	W RR	W ROC	CCI	ICI
PART pruned 75/25 percentage split	94.64	0.06	0.946	0.06	0.947	0.946	0.977	1183	67
PART pruned 10-fold cross validation	**94.66**	0.13	**0.947**	0.059	**0.947**	**0.947**	**0.978**	4733	267

Finally scenario #2 better than scenario #1 model in terms of accuracy, correctly classified instance and ROC curve.

Experimentation V: PART with UnPruned

Table .6. Summary of PART Experiments with unpruned

Experiment	Accuracy	Time Taken	W TPR	W FPR	W PR	W RR	W ROC	CCI	ICI
PART unpruned 75/25 percentage split	93.68	0.27	0.937	0.068	0.937	0.937	0.976	1171	79
PART unpruned 10-fold cross validation	**94.02**	0.28	0.94	0.065	0.94	0.94	**0.984**	4701	299

The last compares is made between Weighted Average ROC curve rate which registered in experiment 4 performances of 97.6% and 98.4% in percentage split and 10- fold cross validation respectively. Finally scenario #2 better than scenario #1 model in terms of accuracy, correctly classified instance, time and ROC curve.

Model Comparison

In this research work, several experiments had been carried out with three classification algorithms, i.e. J48 decision tree algorithm, Naive Bayes classifier and the PART algorithm to build a predictive model that predicts the tumor states in breast cancer Dataset. From the experiments all attributes were identified to make sound rule and better accuracy. Selecting a better classification technique for building a model, which performs best in handling the prediction and identifying significant attribute of tumor states of patient`s is one of the aims of this study.

Table .7. The selected Models Comparison

Selected Model	Accuracy	Precision	Recall	F- measure	Mean absolute error	TPR	FPR	ROC
J48 pruned 10–fold cross validation	**94.82**	**0.948**	**0.948**	**0.948**	**0.0838**	**0.948**	**0.058**	**0.98**
Naive Bayes 10–fold cross validation	87.42	0.877	0.874	0.873	0.2627	0.874	0.147	0.918
PART pruned 10–fold cross validation	94.66	0.947	0.947	0.946	0.0834	0.947	0.059	0.978

Finally, the accuracy achieved on selected feature was 87.42%, 94.66%, 94.82% for Naïve Bayes, PART and J48, respectively.

Generating Rules from Decision Tree

After successive experiments were build and the best decision tree model was selected. The next steps were to generate rules by tracing through the branches up to leafs. The model developed by J48 classifier was selected as the best model for this paper. The generated rules were evaluated by the domain expert. The domain expert agreed on the relevance of the rules, but suggested that further analysis should be performed. The domain expert selected 20 rules that used to develop prototype. From selected one for the purpose explanations of the rule 12 that cover most of the instance in the dataset and high probability to predict (above 80%) are selected by domain expert.

c) Prototype development

The final objective of this study was developing a prototype interface that assists physician easy access to the identified knowledgebase. The final selected if-then rules are used to implement the selected best models.

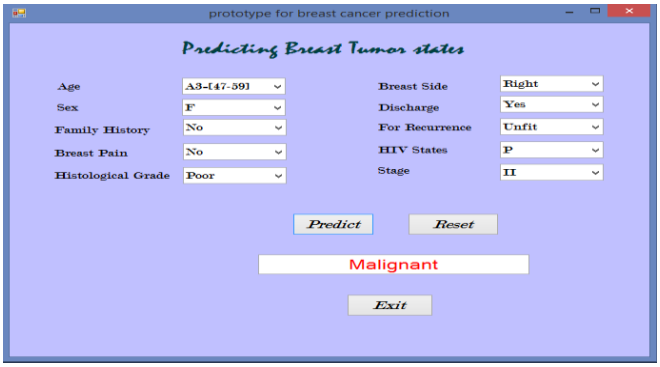


Figure .3. prototypes graphic user interface (GUI)

VI. CONCLUSION

In this paper, the aim was to design a predictive model for breast cancer detection using data mining techniques from breast cancer dataset that is capable of enhancing the reliability of breast cancer disease detection. Hybrid data mining methodology basically follows an iterative process consists of: Business understanding, Data understanding, Data preparation, model building, evaluation and use discover knowledge. The most effective model to predict patients with breast cancer disease appears to be a J48 pruned classifier implemented on 10-Fold Cross Validation with a classification accuracy of 94.82% and still much remains to fill the gap of 5.12% misclassified cases. This means the selected model can also predict breast cancer correctly malignant as malignant or vice versa wrongly with a rate of 5.18%.This has its own implication in reality. Misclassifying malignant as benign means leaving infected person to transmit the disease where as that of benign as malignant is adding tension to patients. Finally, prototype interface develop are developed and the performance of the system is evaluated by the potential users of the system and achieved 82.77% performance.

VII. REFERENCE

[1]. Pharmaceuticals, A., (2010), findings from first-ever initiative on sustainable breast cancer treatment in the developing world, *International Journal of Cancer*,vol.2, pp.134-145

[2]. Fabregue, M, (2011), Mining Microarray Data to Predict the Histological Grade of A *Breast* Cancer *Journal of Biomedical Informatics*, vol. 2 Pp. 4412–4416

[3]. Verma, K., &Zakos, J. (2000), A Computer-Aided Diagnosis System for Digital Mammograms Based on Fuzzy-Neural and Feature Extraction Techniques, *IEEE Transactions on Information Technology in Biomedicine*, vol. 34, pp. 219–223.

[4]. Gupta, S., Kumar D. And Sharma A., (2011), Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis, *Indian Journal of Computer Science and Engineering (IJCSE)*, Vol. 188, pp. 0976-5166

[5]. Ethiopin Ladies, (2014), Available from: *http://www. ethiopianlady.com*

[6]. Han, J. and Kamber M. (2006), Data Mining: Concepts and Techniques, *Amsterdam: Morgan Kaufmann, vol.2,*

[7]. Kaya, Y., (2013), A New Intelligent Classifier for Breast Cancer Diagnosis Based on a Rough Set and Extreme Learning Machine: *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 21, Pp, 2079-2091

[8]. Battling Breast Cancer in Ethiopia, (2015). Available from: http://www.einstein.yu.edu/feature stories/ 1069/ addressing -breast-cancer-in Ethiopia/# sthash.gamh 1P7M. dpuf, 4/2/2016

[9]. Cios, K. and Kurgan, L., (2005), Trends in Data Mining and Knowledge Discovery in Advanced Techniques in Knowledge Discovery and Data Mining, *London: Springer*, Pp. 1–26.