Application of Data Mining Techniques in Early Detection of Breast Cancer

F.Leenavinmalar
Ph.D scholar,Department of computerscience,
Chikkanna govt arts college,Tirupur.
Dr.A.Kumarkombaiya
Assistant professor,Department of computer science
Chikkanna govt arts college,Tirupur

Abstract:*Cancer is a class of diseases characterized by out-of-control cell growth. There are over 100 different types of cancer, and each is classified by the type of cell that is initially affected[1]. Breast cancer is the most common invasive cancer in women, and the second main cause of cancer death in women, after lung cancer, early detection of cancer helps the patients to prevent from vulnerability and get cured. Data mining and machine learning technique are used widely in medical sciences in identifying, diagnosing, diseases. In this paper we are proposing possible data mining techniques in early detection of breast cancer, Wisconsin breast cancer data set is used for experiments, and are evaluated using sensitivity, specificity andclassification accuracy.*

Keywords: *Breast cancer survivability, data mining, Wisconsin breast cancer data set, SVM, C5.0, cancer prediction techniques.*

I. INTRODUCTION

Among different cancers, breast cancer is the very common cancer affecting females worldwide, representing 25% of all cancers,
estimated in 2017 among 1.67 million cancer cases diagnosed. Especially women from less developed countries have slightly more number of cases compared with developed nations (883000 cases against 794000 in developed nations)[1]. In India breast cancer rate is lower (28.8 per 100000) than compared with United Kingdom (95 per 100000) but mortality is at par (12.7 vs17.1 per 100000) with United Kingdom. There is an significant increase in the incident and cancer associated morbidity and mortality in India as global and Indian studies describe.
Earlier cervical cancer was most common cancer in India women but now the breast cancer has surpassed the previous and this leads to increase cancer death. Detecting cancer in its initial stage helps the patients from obtaining medical assistance earlier and avoid death.
In this paper, I have presented the data mining techniques to predict breast cancer valuable and survivability rate of breast cancer patients through wisconsin breast cancer data and SEER[3] data and also introduced a primary classification of data through Survival time recode (STR), Vital status recode (VSR) and cause of death. The next session reviews about the related works, third section gives methodology covers prediction analysis, last session provides the experimental results with conclusion and future works.

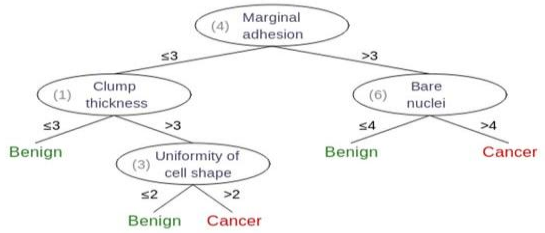
II. RELATED WORKS

A study on the existing techniques and approaches show that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. But we can able to find only few studies related to cancer diagnosis and recurrence using data mining approaches such as C4.5 decision trees. C4.5 is a popular and notable choice tree decision making system which has been utilized by Abdel ghani Bellaachia and Erhan Gauven[9] with two different strategies i.e.

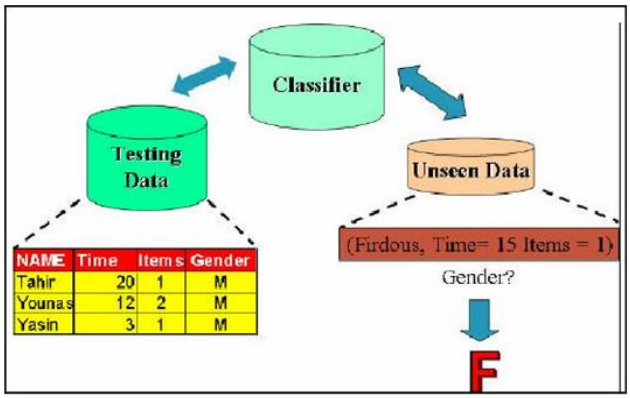
naive Bayes and Back-Propagated Neural Network. The authors examined the expectation of patients survivability rate of breast growth, with help of utilizing SEER Breast Cancer data. SEER data is a preprocessed information set of 151,886 records. With an alternative approach the authors pre-grouped by including Survival Time recode, Vital Status recode and causes of death. They have used Weka tool box to explore data with respect to above three information. With a few investigation led by these calculations, they accomplished to prove these techniques are practically identical to existing procedures. Also mode developed using c4.5 calculation have improved execution that the other two methods.
Authors Wei-stick Chang, Liou and De-ming [3], exhibited that the hereditary calculation of display provided positive results over other information mining models for prediction of breast cancer patients, through the general exactness of the patients, the expression and quality of classification algorithm. The hereditary calculation, decision trees, calculated relapse and

artificial neural system were used for the similar studies and for the positive prescient estimation of every calculation were utilized as the assessment pointers. WBC database was also used for the investigation of data took after by the 10-overlap cross-approval. The results demonstrated that hereditary algorithm delivers exact results in breast cancer disease and the demonstration was adequate, intelligible.

We also reviewed the work of Delen et al[7] the author used a preprocessed the SEER data (period of 1973-2000 with 433,272 records named as breast.txt), for removing redundancies and missing information in breast cancer information to find the survived data accuracy. Which resulted in filtering the data set to 202,932 records, then these data were classified to two groups of 93,273 (Survived) and 109,659 (not survived) these were based on STR filed. Based on these records data mining algorithms were applied to predict the survivability, that resulted in predicting the survivability were in the range of 93%. In this study, the authors proved that data mining could be a valuable tool for predicting breast cancer cases, that can be used for



diagnosis, treatment and prognosis purpose.



Above studies are best examples of applying data mining techniques for breast cancer predictions and also for other medical evaluations.

III. Methodology

In our study we have used SVM, back-propagated neural network and C5.0

Fig. 1.0 C5.0 Decision Tree

decision tree algorithms to investigate the predictability of survivability rate using WCB and SEER breast cancer preprocessed data set. We also found that these three classification techniques most suitable methods for cancer survivability rate prediction.

Support vector machines (SVM)[9] are widely used learning models with associated learning schemes, that is used for classification and regression analysis. With given set of training examples, each set is marked belonging to one or the other category, an SVM scheme assigns new examples to one category or other, and making it a non probabilistic binary liner classifier. SVM is also a representation of examples as points mapped so that each example is separates the categorical and divided by a clear gap that is wide possible.

The other technique uses artificial neural networks, in the technique a multi-layer network with back propagation is used. The third methodology is C5.0 decision tree, this forms a set of training data and C5.0 is based on ID3 algorithm. Through investigating different models and processes we can identify that SVM and C5.0 decision tree techniques perform better in prediction breast tumor and is these techniques are very useful in medical examination practically.

frame-col-ypos:0.593750in;bot-style:none;xpos:0.125694in;wrap-mode:wrapped-both;frame-type:image;frame-page-xpos:4.500694in;ypos:0.039583in;frame-height:1.971528in;frame-pref-page:2;top-style:none;position-to:block-above-text;frame-col-xpos:0.125694in;left-style:none;frame-pref-column:1;right-style:none;frame-width:2.930556in; frame-page-ypos:1.381250in

With our analysis classification predicts categorical class labels (discrete or nominal) & also classifies data (constructs a model) with the training set and the values (class labels) in a classifying attribute is used in classifying new data. The figure 1.0 shows classification of new data.frame-col-ypos:5.786806in; bot-style:none; xpos:0.151389in;frame-width:3.256944in;wrap-mode:wrapped-both;frame-type:image;frame-page-xpos:4.526389in;ypos:1.907639in;frame-height:2.234028in;frame-page-ypos:6.574306in;frame-pref-page:2;position-to:block-above-text;left-style:none;frame-pref-column:1;right-style:none;top-style:none;frame-col-xpos:0.151389in.

Different approaches were adopted in the pre-classification process and we have included STR,VSR and COD. STR field starts from 0 to 180 months in the SEER database and WCB database. This process is outlined as follows

// Setting the survivability dependent variable for 60
// months threshold
if STR ≥ 60 months and VSR is alive then

the record is pre-classified as "survived"
else if STR < 60 months and COD is breast cancer, then
the record is pre-classified as "not survived"
else
Ignore the record
end if

with the above step, we have retrieved the classified the COD is breast cancer or other reasons, and a common analysis may determine the effect the attributes for the prediction and attribute selection. Information gain measure is used to rank the attributes due to the C5.0 decision tree utilize this method.

Information gain (IG) is measured as the amount of the entropy (H) of the difference when an attribute contributes to the additional information about the class. Following is the information gain and the entropy before and after attribute X i is observed for the class C: matrix this can be easily converted to true-positive (TP) and false-positive (FP) metrics

$$H(C) = -\sum p(c)\log p(c)$$
$$H(C|X\,i) = -\sum p(x)\sum p(c|x)\log p(c|x)$$
$$IG\,i = H(C) - H(C|X\,i),\ c \in C,\ x \in X\,i,\ c \in C$$

The study results have be displayed in below tables

Classification Techniques	Accuracy %	Class	Precision	Result
SVM	86.5	0	0.83	0.57
		1	0.87	0.93
Artificial Neural Net	84.5	0	0.7	0.52
		1	0.88	0.97
C5.0	86.7	0	0.8	0.56
		1	0.88	0.96

Table 1 Combined results

Classification Techniques	Accuracy %	Class	Precision	Result
C5.0	81.3	0	0.86	0.81
		1	0.87	0.81

Table 2 C5.0 Results

With above study it shows that C5.0 decision tree algorithm and SVM algorithms are best for classification can cancer data set and can be used for breast cancer predictions. And these decision tree methods are predictive machine learning models decides the targeted value based on various attribute values and available data. The decision tree attributes help in predicting values of the dependent variables in the dataset.

V. **Conclusion**

As discussed in the introduction breast cancer is major cancer disease that affects women and leads to death. The best way to avoid death due this diseases is by predicting it earlier and taking medical steps. Recent technologies and data mining technique can be very useful in predicting breast cancer in earlier stage. In this paper we have listed and demonstrate different data mining techniques that is suitable for predicting the disease.

REFERENCES:

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/)
[2] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, " An Efficient Prediction of Breast
Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013.
[3] Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancerwisconsin/breast-cancer-wisconsin.
[4] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
[5] Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet. 360:1131–1135, 2002.
[6] Witten H.I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.
[7] D. Delen, G. Walker and A. Kadam (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, vol.34, pp.113-127.
[8] Y Rejani- "Early detection of breast cancerusing SVM". 2009 –arxiv
[9] Ilias Maglogiannis, E Zafiropoulos "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers" Applied Intelligence, 2009 –Springer.
[10] Ian H. Witten and Eibe Frank. Data Mining:Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.