Breast Cancer Prediction using Different Classification Algorithm and their Performance Analysis

Koushik Chandra Howlader[1], Urmi Das[2], Mahmudur Rahman[3]
[1]Department Of Computer Science and Telecommunication Engineering
Noakhali Science and Technology University
Noakhali – 3814, Bangladesh
[2]Department of Botany
University of Rajshahi
Rajshahi-6205, Bangladesh
[3]Department of Public Health and Informatics
Jahangirnagar University
Savar, Bangladesh

Abstract: Cancer is a significant health burden worldwide. It is estimated that over 9 million cancer patients are assumed to die in developing countries from different types of cancers by 2030,. The incidence of different types of cancers is increasing due to an unhealthy lifestyle. The main objectives of this paper are to develop a robust data analytical model using various data mining approaches that have been utilized to predict the risk factor of breast cancer and categorized a patient. We have compared the accuracy level of different data mining learning algorithms, and the best model has implemented for predicting disease. Several data mining approaches like a Decision tree, K-Star algorithm, Bayes theorem, Decision Table were used in Weka to find out the most accurate result. The findings of this study may provide a helpful clue to essential facts and figures for breast cancer.

Keywords: Breast Cancer, Prediction, Data Mining, Classification, Neural Network, Naïve Bayes, Decision Tree.

1. Introduction

Data mining has become popular in many organizations because of its powerful intensive and extensive applications and even in healthcare, those applications play vital rule. Its applications can widely help all sections included in the healthcare industry. For extracting medical knowledge, data mining techniques are extremely useful for medical education. Physicians can use data mining applications to identify operative treatments so that patients can get better and more reasonable healthcare services. Matching and mapping strategies become so operative in diagnosis with the help of data mining application. Data in the healthcare industry is really complex and enormous. Dealing with a large amount of data is really hard. Data mining provide several types of methodologies and techniques to process a large number of data and pulling out useful information for decision making which is a fundamental part of the healthcare sector. Experts believe that data mining techniques in the healthcare industry will reduce the cost to 30% of overall healthcare spending. Electronic Health Records (EHR) is quickly becoming more common among healthcare facilities. With increased access to a large amount of patient information, medicinal services providers can now upgrade the capability and nature of their associations utilizing data mining[19]. People from developing countries are not educated enough to understand their diseases. Many of them live in the village. So remote monitoring is critical in healthcare sector especially for those groups' lives far from the city. Because of a considerable number of populations, traditional face-to-face healthcare services emphasize the necessity of remote healthcare service. The study is an effort to apply data mining algorithm on data create a model which can be applied to build a decision support system for physicians and healthcare centers. Prediction models are the core data mining methods mainly used in healthcare and engineering [13-18]. Performance evaluation is done by comparing various models with their accuracy. So we compared our model with the existing model and validated how the proposed model is better than current models.

1.1 Breast Cancer and Its causes

Cancer is defined when abnormal cells are divided in an uncontrolled way. Some types of cancers may eventually spread into other tissues. There exists more than 200 various types of cancer. Cancer develops when gene variations make one cell or a few cells begin to grow and multiply too much. This may cause a growth called a tumor. A primary tumor is the name from where a cancer cell starts to develop. Sometimes cancer can spread to other parts of the body – this is called a secondary tumor or a metastasis. Cancer can affect several body systems, such as the lymphatic and immune systems, the blood circulation system, and the hormone regularity[11]. Most cancers start due to gene changes that happen over a person's lifetime. More rarely cancers start due to inherited faulty genes passed down in families. The majority of cancers, in some of cases 90–95%, are due to genetic mutations from environmental factors[12]. The remaining 5–10% is due to inherited genetics. Various common environmental factors that contribute to cancer death include tobacco (25-30%), infections (15-20%), diet and obesity (30-35%), radiation (both ionizing and non-ionizing, up to 10%), stress, lack of physical activity and pollution. The major symptoms of breast cancer are Abnormal bleeding, Prolonged, Cough, Unexplained weight loss, Change in bowel movements.

2. Literature Review

Comparing Machine Learning algorithms is one of the most common methods for selecting the most appropriate one for a given situation. Many researchers have used this method in evaluating their performance in one or more indicators.

Kalapanidas, et al. [8] worked on the noise sensitivity of 11 ML algorithms (0-Rule, K-NN, Linear Regression, M5, K*, MLP, Decision tables, Hyper pipes, C4.5 Decision trees, C4.5 Rules and Voting Feature Interval). They were based on two prediction problems and one classification problem. They also used four artificial datasets, the first of them develop a multivariate problem, another one is a linear function, while the third and the fourth ones, refer to a non-linear function. All of them had numerical data. Ten-fold cross-validation experiments were carried out for each dataset. K* results were not best but neither the worst for any research, but is essential to note that it is tested here, as a regression algorithm. Even when this study can be considered attractive, they were expressed in graphics, which curves are difficult to be differentiated, so its use could be hard.

The main purpose of Er in [6] was to propose a method for accurate prediction of students assessment in an online course, taking into account log data of the LMS used. All attributes had continuous values (0 - 100), except for one that could be yes or no. Also, they contain a common characteristic: were time-varying. Their evaluated three algorithms: Naïves Bayes, K* and C4.5. The best rates of accuracy, sensitivity, and precision, of the MLalgorithms, were reached for K*. Anyway, this results can only be taken into account with this type of datasets; and just for discerning between this algorithms, according to their performance in this indicators.

Vijayarani and Muthulakshmi [9] compared Lazy (the basic Instance-Based Learner (IBL), k-NN and K*) andBayesian (Bayesian Net and Naïve Bayes) classifiers for text mining problems. They used a dataset of 80000instances and four attributes. Even when lazy classifiers performed better, K* got the highest error rates, and resulted in the worst in the most of the indicators of accuracy (% of correctly/incorrectly classified instances, TP rate, ROC Area and Kappa Statistics). This study offers compelling values of K*, but are not declared details of the kind of values that the dataset used contains.

Douglas et al. in [10] studied on six machine learning algorithms (K*, Naïve Bayes, Support vector classifiers, Decision tree, AdaBoost classification, and Random forest) over a range of complexity; for identifying which is the most accurate in mining functional neuroimaging data. The accuracy of K* was of 86% (the second worst value after Support Vector Machine).

3. Methodology

3.1 Data Collection

Prediction of breast cancer of patients is the primary goal of this paper. So that we had to collect data about breast cancer patients. As there is no data bank about breast cancer in Bangladesh, we have to get data from the Repository. The creator of these data is Matjaz Zwitter & Milan Soklic (physicians), Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia. Those data sets contain nine attributes plus class attribute and 286 Instances.

3.2 Data Preprocessing

In the real world, data are generally incomplete, noisy and inconsistent. For predicting data mining, data in raw form are not best for analysis. Data preprocessing is a first step of the Knowledge discovery in databases (KDD) process. Data preprocessing is a challenging and tedious task. The data must be preprocessed and transformed to get the best mineable form. There are some different tools and methods available for data preprocessing. The tasks in the data preprocessing are: Data Cleaning, Data Integration, Data Transformation, Data Reduction, and Data Discretization. As we liked to do our work in data mining tool Weka[21] we needed to format data to be supported by Weka. To do so, we created an arff file which could be edited by Notepad++.

3.3 Dataset description

The main objective is to forecast if the patient has been affected by cancer using the data mining tools by using the medical data available. The classification type of data mining has been applied to our dataset which had been collected from UCI machine learning repository. Table 3.1 shows a brief description of the dataset that is being considered. There are 13 attribute in our data set in which one attribute is nominal and eleven attribute are numeric and one attribute is class variable. The attributes value distributions are shown in Figure 1 below.

Attributes	Types	value
age	Numeric	10-99
menopause		Lt40,ge40,premeno
tumor-size	Numeric	0-59
inv-nodes	Numeric	0-39
node-caps	Boolean	Yes , no
deg-malig	Numeric	1,2,3
breast		Left,right
breast-quad		Left-up,left-low, right-up,right-low, central
irradiat	Boolean	Yes,no

Table 3.1: Dataset description

3.5 Data analysis and visualization

The distribution of the attribute values with respect to the class attribute is shown in figure1.

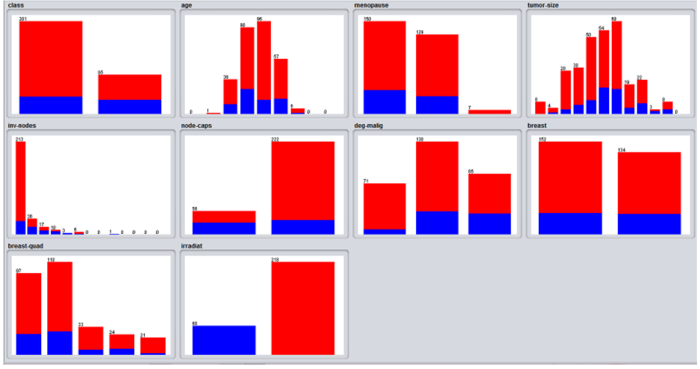

Figure 1 : Attribute value distributions with respect to class attribute

In the picture three colors represents the stage of patients. Blue color stands for the stage "no-recurrence-events" and red color stands for "recurrence-events". In this chapter, different machine learning algorithms and their implementations that have been used for predicting breast cancer are discussed in some details. J48 decision tree, Naïve Bayes, Decision Table, K-Star (Lazy.IBK), Random Tree, LMT Decision Tree, Decision Tree Learning, Logistic Regression [20].

3.4 Proposed Model

Several steps are maintained to build a classification model that analyze and predict the severity of diabetes. First, diabetes patient records are collected that preprocess and extract some features for analyzing further manipulation of it. Then, different classification algorithms are used to classify the severity of diabetes. In the figure we represent several steps how to implement this model. Those steps are described briefly as follows:

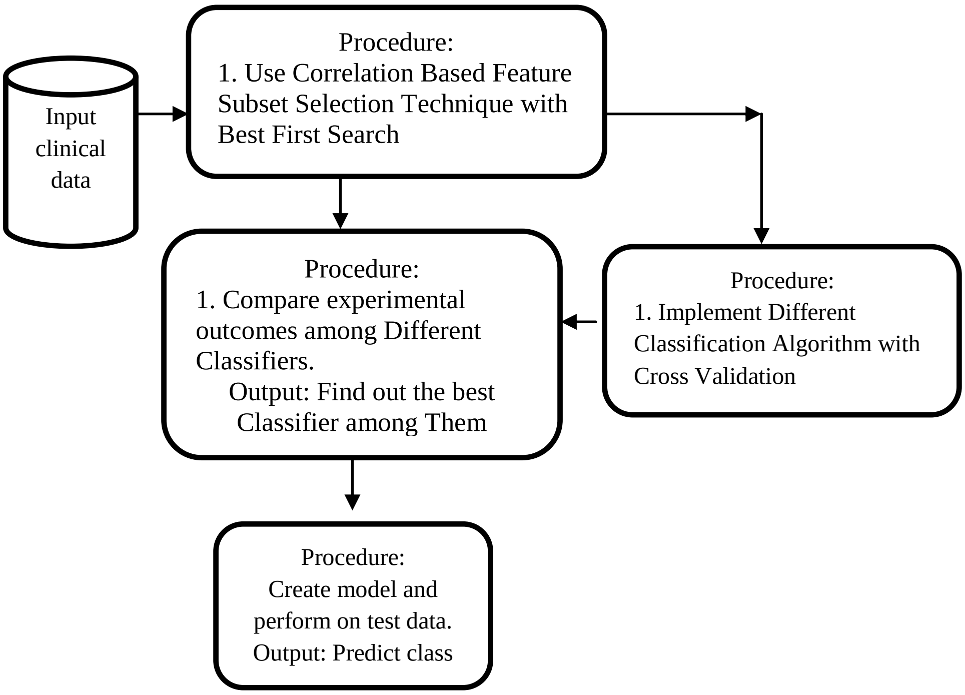
Input clinical data
Procedure:
1. Use Correlation Based Feature Subset Selection Technique with Best First Search
Procedure:
1. Compare experimental outcomes among Different Classifiers.
Output: Find out the best Classifier among Them
Procedure:
1. Implement Different Classification Algorithm with Cross Validation
Procedure:
Create model and perform on test data.
Output: Predict class

Figure 2 Working flow diagram of proposed model

4. Result and Discussion

This chapter discusses in detail, the different experiments that were conducted on the Clinical dataset.

4.1 Model Evaluation with Performance Measure

In our dataset there are two values in a class. Those are ‘recurrence-events, ‘no-recurrence-events’. Based on class values here we created a confusion matrix.

	recurrence-events	no-recurrence-events
no-recurrence-events	a	b
recurrence-events	c	d

Table 4.1: Confusion Matrix

We can evaluate a model by calculating ‘Accuracy’, ‘Precision’, ‘Recall’, ‘F-measure’ and ‘ROC Area’. How these terms can be calculated is described below.

Accuracy $=\frac{a+d}{a+b+c+d}$

Precision(recurrence-events) $=\frac{a}{a+c}$

Precision(no-recurrence-events) $=\frac{d}{d+b}$

True Positive Rate $= \frac{TruePositive}{TruePositive+FalseNegative}$

False Positive Rate $=\frac{FalsePositive}{FalsePositive+TrueNegative}$

Recall (recurrence-events) $=\frac{a}{a+b}$

Recall (no-recurrence-events) $=\frac{d}{d+c}$

F-measure(recurrence-events) $= \frac{2*Precision\,(recurrence-events)*Recall(recurrence-events)}{Precision\,(recurrence-events)+Recall(recurrence-events)}$

F-measure(no-recurrence-events)$=\frac{2*Precision\,(no-recurrence-events)*Recall(no-recurrence-events)}{Precision\,(no-recurrence-events)+Recall(no-recurrence-events)}$

4.1.1 Details of Experiment:

Form different classification algorithm we got a set of their accuracy based on their performance. From table 4.1 we can evaluate the performance of different classification algorithms. Table 4.2 shows the confusion matrix of algorithms for 10 K fold cross validation. Here we found the best accuracy from LMT decision tree and it is 76.2238%. Figure 4.1-4.6 represents the ROC curve of these algorithms.

Classification Algorithm		**No. of Instances**	**Percentage**
LMT decision tree	Correctly Classified Instances	218	76.2238%
	Incorrectly Classified Instances	68	23.7762%
J48 decision tree	Correctly Classified Instances	213	74.4755%
	Incorrectly Classified Instances	73	25.5245%
K-Star	Correctly Classified Instances	215	75.1748%
	Incorrectly Classified Instances	71	24.8252%
Random Tree	Correctly Classified Instances	190	66.4336%
	Incorrectly Classified Instances	96	33.5664%
Decision Table	Correctly Classified Instances	211	73.7762%
	Incorrectly Classified Instances	75	26.2238%
Naïve Bayes	Correctly Classified Instances	210	73.4266%
	Incorrectly Classified Instances	76	26.5734%

Table 4.1: Performance Results from different Classification Algorithm for 10 fold Cross Validation

Table 4.2: Confusion of different Classification Algorithm for 10 fold Cross Validation

Classification Algorithm		recurrence-events	no-recurrence-events
LMT decision tree	no-recurrence-events	191	10
	recurrence-events	58	27
J48 decision tree	no-recurrence-events	191	10
	recurrence-events	63	22
K-Star	no-recurrence-events	183	18
	recurrence-events	53	32
Random Tree	no-recurrence-events	162	49
	recurrence-events	57	28
Decision Table	no-recurrence-events	189	12
	recurrence-events	63	22
Naïve Bayes	no-recurrence-events	170	31
	recurrence-events	45	40

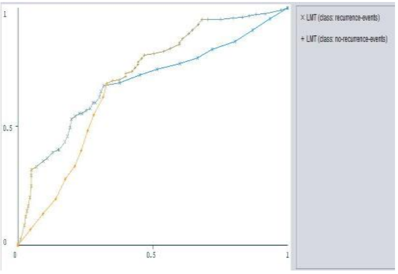

Figure 4.1: Performance Analysis of LMT Classifiers with ROC Curve

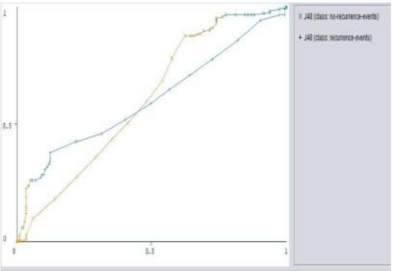

Figure 4.2: Performance Analysis of J48 decision tree Classifiers with ROC Curve

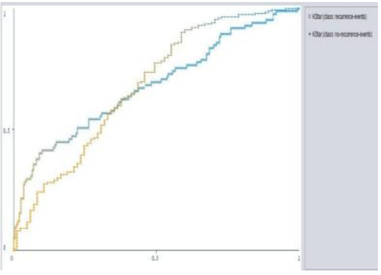

Figure 4.3: Performance Analysis of K Star Classifiers with ROC Curve

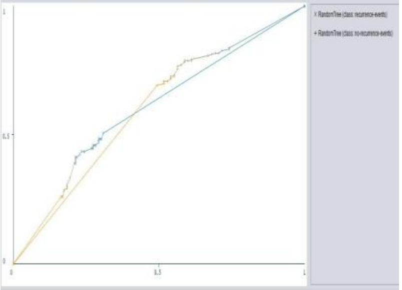

Figure 4.4: Performance Analysis of Random Tree Classifiers with ROC Curve

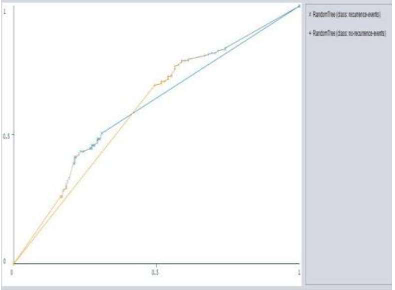

Figure 4.5: Performance Analysis of Decision Table Classifiers with ROC Curve

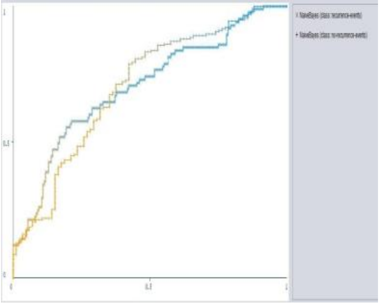

Figure 4.6: Performance Analysis of NaïveBayes Classifiers with ROC Curve

4.2 Result of Classification Algorithms

After performing different classification algorithm in WEKA, value of different term for each classification algorithm is listed in a table. This was done to measure and investigate the performance on the selected classification methods. In this study, all data is considered as instances and features in the data are known as attributes. Here Table 4.3 shows different performance metrics for different classification algorithms. Different performance metrics like TP rate, FP rate, and Precision, Recall, F-measure and ROC area are presented in numeric value. Table 4.4 represents Error measurement for different classifiers in WEKA.

Classifier	**Accuracy**	**Error Rate**	**Kappa Statistic**
LMT Tree	76.2238%	23.7762%	0.32
J48 Tree	74.4755%	25.5245%	0.2549
KStar	75.1748%	24.8252%	0.3256
RandomTree	66.4336%	33.4664%	0.1442
DecisionTable	73.7762%	26.2238%	0.2408
NaiveBayes	73.4266%	26.5734%	0.3321

Table 4.3: Performance measurement for different classifiers in WEKA

Classifier	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root relative squared error**
LMT Tree	0.3547	0.4202	84.7828%	91.9229%
J48 Tree	0.373	0.4406	89.135%	96.3931%
KStar	0.3227	0.4396	77.1363%	96.1702%
RandomTree	0.3411	0.553	81.5133%	120.9891%
DecisionTable	0.3576	0.43	85.4621%	94.0698%
NaiveBayes	0.3237	0.4527	77.3746%	99.0332%

Table 4.4: Error measurement for different classifiers in WEKA

4.3 Model Creation and Prediction

After performing several classification algorithm and ensemble method we saw that shows better result than other classifier. Then we created a model with this ensemble method to make prediction on test data set. We created a test data set where class value was unknown. We used this model to get the class value of test data. After loading the model we reevaluated the model with test data using training set. Most of the prediction is accurate as the main data set.

5. Conclusion and Future Work

Data mining is the procedure of retrieve a pattern from large data set in connection with machine learning, data base and statistics. A data mining technique such clustering, classification and association which is appropriate for medical diagnosis. This paper presents effective classification Techniques. After investigation of different classification Algorithm we have chosen 6 classifier based on our simulation performance and we have used **LMT Tree** classifier achieved overall classification accuracy 76.223% and this algorithm is best this data set. In future we will work to increase this accuracy up to 99% or more.

References

[1]. Tejera Hernández, D. C. (2015). An Experimental Study of K* Algorithm. International Journal of Information Engineering & Electronic Business, 7(2).

[2]. Kalapanidas, E., et al., Machine Learning Algorithms: a Study on Noise Sensitivity. First Balkan Conference in Informatics, 2003 November; pp. 356-365.

[3]. Er, E., Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100.International Journal of Machine Learning and Computing,2012. 2(4): p. 279

[4]. Vijayarani, S.and Muthulakshmi, M., Comparative Analysis of Bayes and Lazy Classification Algorithms.International Journal of Advanced Research in Computer and Communication Engineering, 2013 August; 2(8):3118-3124.

[5]. Douglas, P. K., et al., Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief.Neuroimage. 2011 May 15; 56(2): 544-553.

[6]. A.Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel."Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis." Procedia Computer Science 83 (2016): 1064-1069

[7]. P. Ahmed Iqbal, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab. "Predicting breast cancer recurrence using effective classification and feature selection technique." In Computerand Information Technology (ICCIT), 2016 19th International Conference on, pp. 310-314. IEEE, 2016.

[8]. C. Evandro B., Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, and Joilson Rego. "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses." Computers in Human Behavior 73 (2017): 247-256.

[9]. B. Saba, Usman Qamar, Farhan Hassan Khan, and Lubna Naseem. "HMV: a medical decision support framework using multi-layer classifiers for disease prediction." Journal of Computational Science 13 (2016): 10-25.

[10]. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[11]. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use Data (1973-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2005, based on the November 2004 submission.

[12]. Cox DR. Analysis of survival data. London:Chapman & Hall; 1984.

[13]. Benjamin F. Hankey, et. al. The Surveillance, Epidemiology, and End Results Program: A National Resource. Cancer Epidemiology Biomarkers & Prevention 1999; 8:1117-1121.

[14]. Houston, Andrea L. and Chen, et. al.. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; 13:437-466.

[15]. [15] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; 26:1-24.

[16]. Zhou ZH, Jiang Y. Medical diagnosis with C4.5Rule preceded by artificial neural networkensemble. IEEE Trans Inf Technol Biomed.2003 Mar; 7(1):37-42.

[17]. Lundin M, Lundin J, Burke HB, Toikkanen S,Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999; 57:281-6.

[18]. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[19]. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[20]. J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufmann; 1993.

[21]. Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/