Predicting the Chances of Breast Cancer with Imputation Techniques

Mehta Pankti R.[1], Prof. Mehul Barot[2]
[1, 2] Department of Computer & Information Technology
[1, 2] LDRP-ITR

Abstract- *In modern medical technology, it is various challenges to detect disease at early level. Mainly the cancer.And it is very challenging to give a accurate result if there is missing large number of values in the dataset When it is initial stage it is hard to identify its existence and peoples are die. Many peoples are suffering from cancer mainly women are suffering from breast cancer and many women are died due to the breast cancer. Most of women are illiterate and shay to communicate with her family and husband and also she feel shyness with doctors to check breast cancer. Breast cancer have some symptoms or risk factors which is helpful to detect disease. And the treatment of breast cancer is so expensive. Here we develop a model which is helpful to detect breast cancer at initial stage. For that apply different imputation techniques and replace missing values. Prediction chances of breast cancer predicted using different techniques like K-nearest Neighbor, Multilayer Perceptron, Decision Tree etc. Finally implement the approach for breast cancer which is helpful to predict occurrence chances of breast cancer. Which is very useful for peoples.*

Keywords- Computer science, Clustering, Breast cancer, Tools, Risk factors of Breast Cancer, Women Health Condition.

I. INTRODUCTION

Cancer is a class of most deathful diseases caused by out-of control cell growth. There are more than 100 different types of cancer, and each is classified by the type of cell that is initially affected. Cancer harms the body when damaged cells by dividing uncontrollably. Cancer is treated as most harmful disease in the world because most of the patients died when affected properly. [1]

All of cancers breast cancer is the most in White color women than black color women and 2nd in the world deathful disease. Breast cancer is a cancer that begins in the tissues of the breast. There are two types of Breast cancer one is Ductal carcinoma and another is Lobular carcinoma. Ductal carcinoma begins in the tubes that move milk from the breast to the nipple. Most breast cancers are of this type. It may progress to invasive cancer if untreated. Lobular carcinoma starts in the parts of the breast, called lobules, which produce milk. [1] Breast cancer has 100 times more chances to occur in women than in men. [2].

Over the years the use of various kinds of prediction\classification models in the medical domain has been amplified largely due to their effectiveness and improved prediction capability. Because predicting the accurate disease outcome would help physicians in making less invasive decisions, consequently relieving the patient from having unnecessary adjuvant treatments and their colossal costs. Cancer prediction\prognosis has three directions of prediction analysis: **1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability[2].** This manuscript focuses the survey of research efforts done in the area of prediction of breast cancer survivability. The term "Survival" defines the time period of patient staying alive after the diagnosis of the disease. In most of the research efforts "Survival" is considered as the incidence of breast cancer where individual is still alive after 5 years (sixty months or 1825 days) from the date of diagnosis.[2]

Risk Factors:

Gender, Age, Genetic Risk factors, family history, personal history, Dense breast tissues,Certain benign (not cancer) breast problems, Lobular carcinoma in situ,Menstrual periods, Not having children or having them later in life, Using hormone therapy after menopause, Not breastfeeding, Alcohol, Being overweight or obesity.[3]Treatment for breast cancer depends on the type and stage of the disease, the size the tumor, general health, medical history and age of patients. In most cases, the goal of treatment is to remove or destroy the cancer completely.Like other diseases treatment, breast cancer treatment has also some side effects. It may damages healthy cells and tissues, unwanted side effects sometimes occur. Side effects depend mainly on the type and extent of the treatment. Side effects may not be the same for each person. So it tells again that "Prevention is better than cure". [1]

II. BACKGROUND STUDY

Data mining has some fields to analysis of data such as classification, clustering, correlations, association rule etc. Now-a-days data mining has been used intensively and extensively by many organizations. In Health-care, data mining is becoming increasingly popular. Data mining provides the methodology and technology to analysis the useful information of data for decision making. Pre-processed data are broken down into two categories: relevant and non-relevant dataset to breast Cancer. Clustering technique mainly used to find out two types of data. Clustering is a process of separating dataset into subgroups according to the unique feature. WEKA toolkits are mainly used to find out highest affected risk factors as well as ranking of risk factors from dataset.[1]

Data Mining

Data Mining came into existence in the middle of 1990's and appeared as a powerful tool that is suitable for fetching previously unknown pattern and useful information from huge dataset. Various studies highlighted that Data Mining techniques help the data holder to analyze and discover unsuspected relationship among their data which in turn helpful for making decision.[8]

In general, Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process.[8]

According to Fayyad et al., the knowledge discovery process are structured in various stages whereas the first stage is data selection where data is collected from various sources, the second stage is preprocessing of the selected data , the third stage is the transformation of the data into appropriate format for further processing, the fourth stage is Data Mining where suitable Data Mining technique is applied on the data for extracting valuable information and evaluation is the last stage[8]

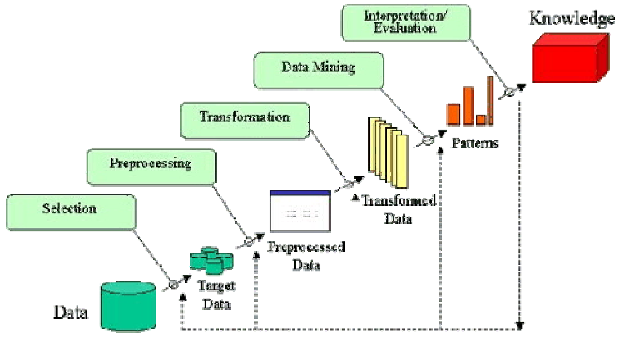


Fig.-1 KDD Process [8]

2.1 Data Mining Methods

As the amount of data stored in medical databases is increasing, there is growing need for efficient and effective techniques to extract the information. Previous researches have given evidence that medical diagnosis and prognosis is amended by employing data mining techniques on clinical data (Hammer & Bonates, 2006; Saastamoinen & Ketola, 2006; Tsirogiannis et al., 2004). This has been possible due to extensive availability of data mining techniques and tools for data analysis. Predictive modeling requires that the medical informatics researchers and practitioners need to select the most appropriate strategy to cope with clinical prediction problem (Bellazzi & Zupan, 2008). [4].

2.1.1 Decision Methods

Classification:

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization.[8]

2.1.1.1 KNN classifiers[8]

When the KNN method is used to classify an unlabeled pattern x, the distances from x to the labeled instances are computed using the HEOM, its K Nearest Neighbors are found, and their class labels are then used to assign a class to x. [5]

This research work used K-NN to analyze the relationship between cardiovascular disease and hypertension and the risk factors of various chronic diseases in order to construct an early warning system to reduce the complication occurrence of these diseases [8]

2.1.1.2 Decision Tree[8]

DT is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node. For example we have a financial institution decision tree which is used to decide that a person must grant the loan or not. Building a decision for any problem doesn't need any type of domain knowledge. Decision Trees is a classifier that use tree-like graph. The most common use of Decision Tree is in operations research analysis for calculating conditional probabilities [8].

Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain. Decision Tree is widely used by many researchers in healthcare field. [8]

2.1.1.3 Support Vector Machines[8]

The concept of SVM is given by Vapnik et al., which is based on statistical learning theory. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems. The support vector machine classifier creates a hyper plane or multiple hyper planes in high dimensional space that is useful for classification, regression and other efficient tasks. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space. Kernel functions are used for non-linear mapping of training samples to high dimensional space. Various kernel function such as polynomial, Gaussian, sigmoid etc., are used for this purpose. SVM works on the principal that data points are classified using a hyper plane which maximizes the separation between data points and the hyper plane is constructed with the help of support vectors [8].

2.1.1.4 Neural Network (NN)[8]

It is an algorithm for classification that uses gradient descent method and based on biological nervous system having multiple interrelated processing elements known as neurons, functioning in unity to solve specific problem. Rules are extracted from the trained **Neural Network (NN)** help to improve interoperability of the learned network [8]. To solve a particular problem NN used neurons which are organized processing elements. Neural Network is used for classification and pattern recognition. An NN is adaptive in nature because it changes its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the information that flows internally and externally through network during learning phase. In NN multiclass, problem may be addressed by using multilayer feed forward technique, in which Neurons have been employed in the output layer rather using one neuron. Er et al., construct a model using Artificial Neural Network (ANN) for analyzing chest diseases and a comparative analysis of chest diseases was performed using multilayer, generalized regression, probabilistic neural networks [8].

2.1.1.5 Multilayer Perceptron (MLP)[6][10]

The multi-layer perceptron (MLP) model is capable of mapping set of input data into a set of appropriate output data. The primary task of neurons in input layer is the division of input signal Xi among neurons in hidden layer. The output of neurons in the output layer is determined in an identical fashion. Figure 2 shows MLP feed forward Neural Network. The back-propagation algorithm can be employed effectively to train neural networks; it is widely recognized for applications to layered feed-forward networks, or multi-layer perceptrons. MLP is the most commonly used algorithm and performs better than other ANN architectures for this type of classification problems. [6]

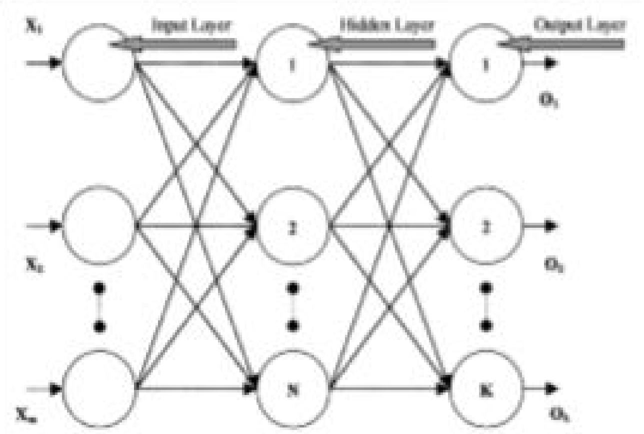


Fig.-2 MLP Feed forward neural network [10]

2.2 Missing value Imputation

2.2.1 Introduction[4]

In real-life databases, incomplete data or sufficient information are not available. Missing values can occur due to large number of reasons such as errors in the manual data entry procedures, equipment errors or incorrect measurements. The presence of missing values (MVs) in data mining produces several problems in the knowledge extraction process such as loss of efficiency, complications in managing and analyzing data. It may also result in bias decisions due to differences between missing and complete data.[4]

In order to solve these problems, two approaches are found in the literature. First approach consists of missing data toleration techniques which integrate the techniques of missing values handling in specific data mining algorithms such as in classification (David, 2007; Saar-Tsechansky, 2007), clustering (Hathaway & Bezdek, 2002) and feature selection (Aussem & de Morais, 2008). Second type of approach consists of missing data imputation techniques which fill in missing values before using complete-data methods on data sets. One advantage of imputation is that the treatment of missing data is independent of the succeeding mining algorithm, and people can select a suitable learning algorithm after imputation.[4]

Missing data randomness can be divided into three classes, as proposed by[22]

1. Missing completely at random (MCAR).

This is the highest level of randomness.

It occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method can applied without risk of introducing bias on the data[22];

2. Missing at random (MAR).

When the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of the missing data itself[22];

3. Not missing at random (NMAR).

When the probability of an instance having a missing value for an attribute could depend on the value of that attribute[22].

2.2.2 Imputation Techniques[4]

There are different Techniques:
Case deletion, Most Common Method(MC), Concept Most Common(CMC), k-Nearest neighbor(KNNI), Weighted Imputation with K-Nearest Neighbor(WKNN), K-means Clustering Imputation(KMI), Imputation with Fuzzy K-means Clustering(FKMI), Support Vector Machines Imputation(SVMI), Singular Value Decomposition Imputation (SVDI), Local Least square Imputation (LLSI), Matrix Factorization.[4]

According to the literature, more appropriate treatments for dealing with MD are the following ones[5]:

- Imputation based approaches: Using the available complete data, the MD are estimated and filled by plausibl evalues.Afterthat,the classifier is designed using the imputed dataset. Therefore, two different and consecutive stages could be considered: imputation and classification.[5]
- Avoiding explicit imputations: In this kind of approaches, the decision methods are able to deal with unknown values, i.e., they can handle MD during the classifier design. Then, classification is performed without a previous MD imputation.[5]

2.3 Performance evaluation[4][5]

In this work, four well-known measures for performance evaluation in prediction problems[4] were employed: accuracy, sensitivity, specificity and area under ROC Curve. The first three measures are given by:

$$\textbf{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad \textbf{(1)}$$

$$\textbf{Sensitivity} = \frac{TP}{(TP+FN)} \qquad \textbf{(2)}$$

$$\textbf{Specificity} = \frac{TN}{(TN+FP)} \qquad \textbf{(3)}$$

Where,
True positives (TP) = No. of correct classifications predicted as yes (or positive).[4]

True negatives (TN) = No. of correct classifications predicted as no (or negative).[4]

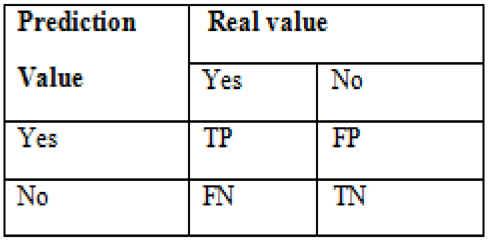
False positive (FP) = No. of examples that are incorrectly predicted as yes (or positive) when it is actually no (negative).[4]

False negative (FN) = No. of examples that are incorrectly predicted as no when it is actually yes.[4]

These four values are summarized in a 2X2 confusion matrix (see Table 1).

In addition to Eqs. (1)–(3), the area under the ROC curve (AUC) is an effective way to measure the overall performance. This measure represents a trade-off between sensitivity and specificity. AUC takes values from 0 to 1, where 0 indicates a perfectly inaccurate model and 1 reflects a perfectly accurate model. In general, a value of 0.5 for AUC is considered as the lower bound. The higher the AUC is, the greater the overall performance of the model to correctly predict survival.[5]

Table-1: Confusion Matrix [5]



Prediction Value	Real value	
	Yes	No
Yes	TP	FP
No	FN	TN

III. GOALS & OBJECTIVES

Goals:

Goal of this work is make awareness to peoples of Breast cancer.

Main goal of this work is make aware the women about cancer and they gain knowledge about breast cancer and also know about symptoms of breast cancer.

This work gives the chances of breast cancer at early stage.

So, the patient became aware for that disease they are know that they are suffer form that disease and take appropriate treatment and care for that.

Using this approach we will reduce the death ratio of women who are died because of breast cancer.

Objectives:

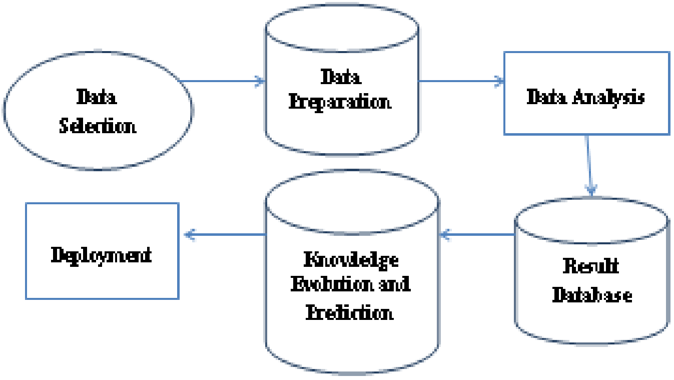


Fig.-3 Objectives

IV. METHODOLOGY

In this work take the database which contains the missing attributes.

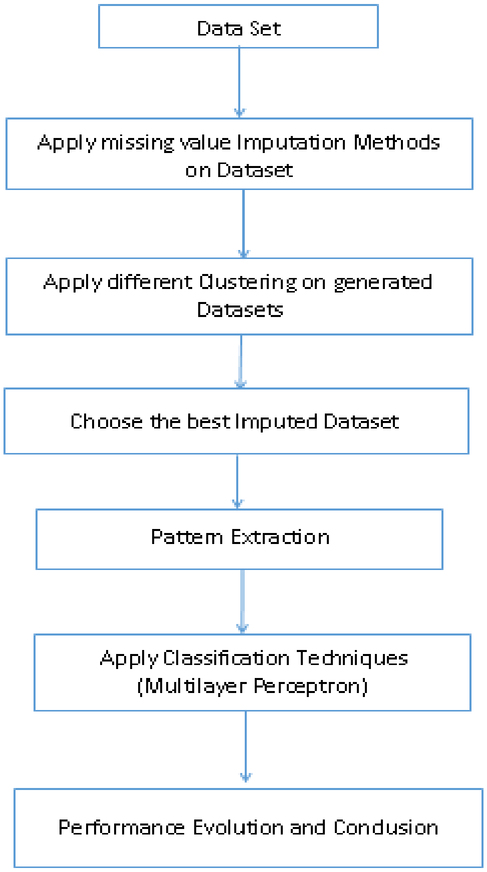

Data Set
Apply missing value Imputation Methods on Dataset
Apply different Clustering on generated Datasets
Choose the best Imputed Dataset
Pattern Extraction
Apply Classification Techniques (Multilayer Perceptron)
Performance Evolution and Conclusion

Fig.-4 Process for Breast cancer
Fig.-4.1 Flowchart for Proposed work

After that apply imputation techniques like Case deletion, Most Common Method(MC), Concept Most Common(CMC), k-Nearest neighbor(KNNI), Weighted Imputation with K-Nearest Neighbor(WKNN), so we get the different data bases (like D1, D2, D3,.., Dn). In this datasets missing values are replaces with some new values. And then apply classifier to all datasets and examine the result of all dataset and finally we get the result. Based on the result we will conclude which imputation technique is better than others.

V. CONCLUSION

Here, we studied different data mining techniques which are used to take decision about health information and we also studied about different missing value techniques and Multilayer Perceptron algorithm.

REFERENCES

[1] "Prediction of Breast Cancer Risk Level with Risk Factors in Perspective to Bangladeshi Women using Data Mining". International Journal of Computer Applications (0975 – 8887) Volume 82 – No4, November 2013.

[2] "A Survey of Prediction Models for Breast Cancer Survivability". ICIS 2009, November 24-26, 2009 Seoul, Korea Copyright © 2009 ACM 978-1-60558-710-3/09/11... $10.00"

[3] "A Novel Approach for Breast Cancer Detection using Data Mining Techniques. " International Journal of Innovative Research in Computer and Communication Engineering. (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014

[4] "Hybrid Prediction Model With missing value imputation for medical data." journal home page: ww.Elsevier.com/locate/eswa. Expert system with application42(2015) 5621-5631

[5] "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values" http://dx.doi.org/10.1016/j.compbiomed.2015.02.006. 0010-4825/& 2015 Elsevier Ltd. All rights reserved.

[6] "Predicting Breast Cancer Recurrence Using Machine Learning Techniques". International Journal of Latest Trends in Engineering and Technology (IJLTET)

[7] "Breast Cancer Prevention and Early Detection " American Cancer Society

[8] "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013),pp.41-266 http://dx.doi.org/1 0.14257/ ijbsbt.2013.5.5.25

[9] "Different Machine Learning Algorithm for Breast cancer Diagnosis." International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.6, November 2012

[10] "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence". Ahmad et al., J Health Med inform-2013,4:2 http://dx.doi.org/10.4172/2157-7420.1000124

[11] "Septic Shock Prediction for Patients with Missing Data". ACM Transactions on Management Information Systems, Vol. 5, No. 1, Article 1, Publication date: April 2014.

[12] "Predicting Breast Cancer Survivability Using Data Mining Techniques". Abdelghani Bellaachia, Erhan Guven . Department of Computer Science The George Washington University Washington DC 20052 {bell, eguven}@gwu.edu

[13] "Robust Predictive model for evaluating breast cancer survivability". Journal homepage: www.Elsevier.com/locate/engappai Engineering Application of Artificial Intelligence 26(2013) 2194-2205

[14] "MED-StyleR: METABO Diabetes-Lifestyle Recommender" RecSys2010, September 26–30, 2010, Barcelona, Spain. Copyright 2010 ACM 978-1-60558-906-0/10/09

[15] "Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges." Int. J. Environ. Res. Public Health 2014, 11, 2580-2607; doi:10.3390/ijerph110302580

[16] "Missing data imputation using statistical and machine learning methods in a real breast cancer problem" Artificial Intelligence in Medicine 50 (2010) 105–115. Journal homepage: www.elsevier.com/locate/eswa

[17] "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS" Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE)

[18] "Early Detection and Prevention of Cancer using Data Mining Techniques " International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014

[19] "An Overview on Data Mining Approach on Breast Cancer data" International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-3 Number-4 Issue-13 December-2013

[20] "Predicting Individual Disease Risk Based on Medical History"CIKM'08, October 26–30, 2008, Napa Valley, California, USA. Copyright 2008 ACM 978-1-59593-991-3/08/10 ...$5.00.

[21] "A breast cancer prediction model incorporating familial and personal risk factors". STATISTICS IN MEDICINE Statist. Med. 2004; 23:1111–1130 (DOI: 10.1002/sim.1668)

[22] "A Study of K-Nearest Neighbour as anImputation Method"Gustavo E. A. P. A. Batista and Maria Carolina Monard University of S˜ao Paulo - USP