In India the event of breast cancer cases are increasing day by day. A new global study estimates that by 2030, the number of new cases of breast cancer in India will increase from the current 115,000 to around 200,000 per year.Cancer treatment and early successful diagnosis of the patients is a challenge since so many years. Doctors and Researchers have been working every day to find new ways to treat cancer.Data mining for cancer treatment can become a great support tool for doctors and physicians for decision making and estimation purpose. The need for biological data mining is that there is too much data but they are mostly unstructured.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Journal Article

% LaTeX Template

% Version 1.3 (9/9/13)

%

% This template has been downloaded from:

% http://www.LaTeXTemplates.com

%

% Original author:

% Frits Wenneker (http://www.howtotex.com)

%

% License:

% CC BY-NC-SA 3.0 (http://creativecommons.org/licenses/by-nc-sa/3.0/)

%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%----------------------------------------------------------------------------------------

% PACKAGES AND OTHER DOCUMENT CONFIGURATIONS

%----------------------------------------------------------------------------------------

\documentclass[twoside]{article}

\usepackage{lipsum} % Package to generate dummy text throughout this template

\usepackage[sc]{mathpazo} % Use the Palatino font

\usepackage[T1]{fontenc} % Use 8-bit encoding that has 256 glyphs

\usepackage{algorithm}

\usepackage{graphics}

\usepackage{setspace}

\usepackage{epsfig}

\usepackage{graphicx}

\linespread{1.05} % Line spacing - Palatino needs more space between lines

\usepackage{microtype} % Slightly tweak font spacing for aesthetics

\usepackage[hmarginratio=1:1,top=32mm,columnsep=20pt]{geometry} % Document margins

\usepackage{multicol} % Used for the two-column layout of the document

\usepackage[hang, small,labelfont=bf,up,textfont=it,up]{caption} % Custom captions under/above floats in tables or figures

\usepackage{booktabs} % Horizontal rules in tables

\usepackage{float} % Required for tables and figures in the multi-column environment - they need to be placed in specific locations with the [H] (e.g. \begin{table}[H])

\usepackage{hyperref} % For hyperlinks in the PDF

\usepackage{lettrine} % The lettrine is the first enlarged letter at the beginning of the text

\usepackage{paralist} % Used for the compactitem environment which makes bullet points with less space between them

\usepackage{abstract} % Allows abstract customization

\renewcommand{\abstractnamefont}{\normalfont\bfseries} % Set the "Abstract" text to bold

\renewcommand{\abstracttextfont}{\normalfont\small\itshape} % Set the abstract itself to small italic text

\usepackage{titlesec} % Allows customization of titles

\renewcommand\thesection{\Roman{section}} % Roman numerals for the sections

\renewcommand\thesubsection{\Roman{subsection}} % Roman numerals for subsections

\titleformat{\section}[block]{\large\scshape\centering}{\thesection.}{1em}{} % Change the look of the section titles

\titleformat{\subsection}[block]{\large}{\thesubsection.}{1em}{} % Change the look of the section titles

\usepackage{fancyhdr} % Headers and footers

\pagestyle{fancy} % All pages have headers and footers

\fancyhead{} % Blank out the default header

\fancyfoot{} % Blank out the default footer

\fancyhead[C]{Techno India University} % Custom header text

\fancyfoot[RO,LE]{\thepage} % Custom footer text

%----------------------------------------------------------------------------------------

% TITLE SECTION

%----------------------------------------------------------------------------------------

\title{\vspace{-15mm}\fontsize{24pt}{10pt}\selectfont\textbf{A Study on Risk Factors \& Cell Behaviour of Breast Cancer}} % Article title

\author{

\large

\textsc{Supratim Bhattacharya\thanks {Corresponding Author}\space\space\space,\space Jayanta Porey \& Sampurna Mandal}\\[2mm] % Your name

Department of Computer Science \& Engineering, Techno India University,\\Saltlake, kolkata - 700091, India\\

\normalsize \href{mailto:bhattacharya.supratim@gmail.com}{bhattacharya.supratim@gmail.com \& jayanta.poray@gmail.com} % Your email address

\vspace{-5mm}

}

\date{}

%----------------------------------------------------------------------------------------

\begin{document}

\maketitle % Insert title

\thispagestyle{fancy} % All pages have headers and footers

%----------------------------------------------------------------------------------------

% ABSTRACT

%----------------------------------------------------------------------------------------

\begin{abstract}

Women's most threatened and deadliest diseases is breast cancer. It is the second leading cause to death for women today and it is the most common cancer in developed countries, accounting for >1.6 \% of deaths and case fatality rates are highest in low-resource countries. There are some modifiable(BMI, age at first child birth, number of children, duration of breast feeding, alcohol, diet and number of abortions ) and non-modifiable(age, gender, number of first degree relatives suffering from breast cancer, menstrual history, age at menarche and age at menopause) factors associated with risk assessment of breast cancer.

\vspace{.5cm}

As the level of recurrence for breast cancer is high, it is important to do a proper diagnosis. Number and size of medical databases are increasing rapidly, so we can able to extract hidden knowledge from this data. Advanced data mining techniques can be used to discover hidden patterns and relationships. Models developed from these techniques are useful for medical practitioners to make right decisions. Our research studied the data, applied on WEKA(Waikato Environment for Knowledge Analysis), as a data mining tool to develop predictive models for breast cancer recurrence patients from the data available in Winconsin database .

\end{abstract}

%----------------------------------------------------------------------------------------

% ARTICLE CONTENTS

%----------------------------------------------------------------------------------------

\begin{multicols}{2} % Two-column layout throughout the main article text

**\section{Introduction}**

\lettrine[nindent=0em,lines=3]{I}n India, the average age of developing a breast cancer , has undergone a significant shift over the last decades. According to the official Indian Registries, subset of the National Cancer Registry Programme, women ages between 30 to 50 years has a significant increase in positive malignancy compare to 25 years back. However the age between 40 to 60 years is of high threatening for women. Researches estimated that 636,000 cases occured in developed countries and 514,000 in developing countries during 2002\cite{Nassif:$2009$}.Currently a women living in US has a 12.3\% lifetime risk of developing breast cancer \cite{Nassif:$2009$}. It has been seen that $\frac{1}{4}^{th}$ of all female cancer is breast cancer. Breast cancer is now the most common cancer in most cities in India and $2^{nd}$ most common on the real areas and India is a country with largest estimated number of breast cancer.

Breast cancer is one of the most common cancers among Egyptian women; as it represents $18.3$ \% of the total general of cancer cases in Egypt and a percentage of $37.3$ \% of breast cancer is considered treatable disease. Early diagnosis helps to save thousands of disease victims. The age of breast cancer affection in Egypt and Arab countries is prior ten years compared to foreign countries as the disease targets women in the age of $30$ in Arab countries, while affecting women above $45$ years in European countries. Breast cancer comes in the top of cancer list in Egypt by $42$ cases per $100$ thousand of the population. However $80$ \% of the cases of breast cancer in Egypt are of the benign kind\cite{Elattar}. The industrialized nations such as the United States, Australia, and countries in Western Europe witnessed the highest incidence rates. In many countries, breast cancer incidences increased during the 20th century, largely reflecting global changes in reproductive patterns and regional increases in mammography\cite{Daniel}. Because of social and cultural considerations, breast cancer ranks highest among women's health concerns. It is the most frequently diagnosed cancer in women. After thyroid cancer, melanoma, and lymphoma, breast cancer comes fourth in cancer incidences in women between $20$ to $29$ years. Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy.

Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these approaches to classification and prediction systems, especially in helping medical practitioners in their decision making. In addition to its importance in finding ways to improve patient outcomes, reduce the cost of medicine, and help in enhancing clinical studies. Although there was a great deal of public education and scientific research, Breast cancer considered the most common invasive cancer in women, with more than one million cases and nearly 600,000 deaths occurring worldwide annually\cite{Lyon IAfRoC}

Breast cancer affecting about 10\% of all women at some stages of their life. In recent years, the incidence rate keeps increasing and data shows that the survival rate is 88\% after five years from diagnosis and 80\% after 10 years from diagnosis.

Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy.

We have applied WEKA machine learning technique to analyze the risk factors so that early diagnosis and proper preventive measures can be taken. We have used the Wisconsin Diagnostic Breast Cancer Dataset from UC Irvine for our study. The classification is based on benign or malignant.

%------------------------------------------------

**\section{Problem Definition}**

The dataset used in this experiment were obtained from Wisconsin Diagnostic Breast Cancer Dataset from UC Irvine machine learning repository and described by Dr. William H. Wolberg. The breast cancer data have been used in some other research. We study the effect of nine characteristics parameter on the state of Breast cancer and the influence of the involved parameters on the performance of the SVM model. We have used WEKA Tool as a classifier in our experiment. Our aim is to predict the various state, behaviour and characteristics of breast cancer.

In this dataset, there are 698 samples taken from different women and every sample is expressed by nine characteristic parameter. The nine parameter are as follows:- Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, single epithelia cell size, Bare Nuclei, Bland chromatic, Normal Nucleoli, Mitoses. According to the properties of these nine parameter, the brest cancer is classified into benign \& malignant. Every single parameter is given a range between 1 to 10 and the resultant class is expressed by 2 for benign and 4 for malignant. Among the 698 samples in the dataset there are 16 samples with missing or incomplete data. So we have used remaining 682 samples in this machine learning.

**\section{Proposed Model}**

Fig 1 shows the functional block diagram of our proposed model. It consists of four steps: (a) Acquisition, (b) Preprocess, (c) Feature Extraction and (d) Feature Selection.

In acquisition step, feature selection \& feature extraction is accomplished in order to determine the input vector and based on either feature selection or feature extraction, dimensionality reduction is accomplished. In the preprocessing phase, filtering is done to clear the noise \& map the entire data into lower dimension. Also less important and redundant information are ignored. In the classification step different classifier is used to get the best result out of it. We have also applied clustering method \& Association Rule Mining to obtain more decisions \& to predict more accurately.

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=0.2]{fig1.png}

\caption{Proposed Model}

\end{center}

\end{figure}

**\section{Dataset Description}**

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used \cite{Frank}, to distinguish malignant (cancerous) from benign (non-cancerous) samples. A brief description of these datasets is presented in table 1.Each dataset consists of some classification patterns or instances with a set of numerical features or attributes.

\vspace{.5 cm}

\begin{center}

\textbf{Table1: DESCRIPTION OF BREAST CANCER DATASETS}

\end{center}

\begin{tabular}{|p{1.5 cm}|c|c|c|}

\hline

Dataset & Attribute & Instant & Class\\

\hline

Wisconsin BC & 11 & 698 & 2\\

\hline

\end{tabular}

**\section{Methodology}**

The dataset's attributes are found listed in Table 2.

\vspace{.5 cm}

\begin{center}

\textbf{Table2: WISCONSIN BREAST CANCER DATASET ATTRIBUTES}

\end{center}

\begin{tabular}{|c|c|c|}

\hline

& Attribute & Domain\\

\hline

1 & Sample Code No & id no\\

\hline

2 & Clumb Thickness & 1-10\\

\hline

3 & Uniformity(Cell Size) & 1-10\\

\hline

4 & Uniformity(Cell Shape) & 1-10\\

\hline

5 & Marginal Adhesion & 1-10\\

\hline

6 & sgl Epithelial(cell size) & 1-10\\

\hline

7 & Bare Nucleoli & 1-10\\

\hline

8 & Normal Nucleoli & 1-10\\

\hline

9 & Mitoses & 1-10\\

\hline

10 & Class & 2 or 4\\

\hline

\end{tabular}

\vspace{.5 cm}

In the Clump thickness benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in

multilayered. While in the Uniformity of cell size/shape the cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not. In the case of Marginal adhesion the normal cells tend to stick together, where cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy. In the Single epithelial cell size the size is related to the uniformity

mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell. The Bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. The Bland Chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser. The Normal nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible. In cancer cells the nucleoli become more prominent, and

sometimes there are more of them. Finally, Mitoses is nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses.

We next present our algorithm and further describe the dataset on which we have evaluated.

our first step is to discretize the dataset into three major groups.

\begin{enumerate}

\item Low

\item Mid

\item High

\end{enumerate}

\begin{algorithm}[H]

\caption{An algorithm for discretization of dataset}

\textbf{Input} : Dataset in excel format with $9$ parameters.

\textbf{Output:} : Dataset in csv file(space delimiter) format in discrete format with all $9$ parameters.

\textbf{Algorithmic Steps:}

\begin{enumerate}

\item Obtain the ranges of high, middle and low.

\item collect every cell value for computation for every parameter.

\item \textbf{For} parameter $1$ to $9$ do

\item \textbf{If} Cell value>= high value \textbf{Then}

\textbf{Put} new Cell value= 'H'

\textbf{Else If} Cell value>= middle value \textbf{Then}

\textbf{Put} new Cell value= 'M'

\textbf{Else}

\textbf{Put} new Cell value= 'L'

\item \textbf{End if}

\item \textbf{Next}

\item \textbf{For} $10^th$ parameter

\item \textbf{If}Cell value=2 \textbf{Then}

\item \textbf{Put} new Cell value = "Benign"

\item \textbf{Else If}Cell value=4 then

\item \textbf{Put}new Cell value = "Malignant"

\item \textbf{End If}

\item Construct another excel file based on this discrete value.

\item Convert the excel file into csv(space delimiter) file.

\end{enumerate}

\end{algorithm}

As the parameter of the dataset ranges from 1 to 10 we made this discretization based on different ranges like:-

\begin{enumerate}

\item low:- 1 to 1 mid:- 2 to 6 high:- 7 to 10

\item low:- 1 to 1 mid:- 2 to 7 high:- 8 to 10

\item low:- 1 to 2 mid:- 3 to 7 high:- 8 to 10

\item low:- 1 to 3 mid:- 4 to 6 high:- 7 to 10

\item low:- 1 to 4 mid:- 5 to 6 high:- 7 to 10

\item low:- 1 to 4 mid:- 5 to 7 high:- 8 to 10

\end{enumerate}

%------------------------------------------------

**\section{Analysis \& Results}**

We have applied this modified dataset in WEKA Tool for further analysis and we got the following result:-

Observations:-

\begin{enumerate}

\item We observe that the value of Clump Thickness \& Bare Nucleoli tends to be in higher side. More than $20\%$ of the values are in higher side, compare to other parameters who ranges on $16\%$ in higher side.

\item More than $35\%$ of the value for the parameter Clump thickness, Epithelial size \& bland chromatin ranges in medium side.

\item Bare Nucleoli's medium range value is in $< 11\%$ data whereas others had an average of $20\%$.

\item Malignancy is positive when Clump Thickness \& Bare Nucleoli is in higher side but Size Uniformity \& Shape Uniformity has a very sensitive effect, i.e when the value is $>=4$ there shows a positive sign of malignancy in more than $80\%$ cases.

\item More than $64\%$ malignancy is positive only due to these two parameters.

\item For marginal adhesion \& epithelial size, value ranges between $2$ to $5$. More than $40\%$ cases it is malignancy when this two parameter is low.

\item In case of Bare Nucleoli \& Normal Nucleoli, in $70\%$ cases they tends to be low. They also show positive malignancy in $40\%$ cases when they are low. They have $60\%$ values in between $3$ to $4$. If the value is $5$ or more then definitely it is malignancy.

\end{enumerate}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.7]{fig2.png}

\caption{J48 pruned tree}

\end{center}

\end{figure}

\vspace{1 cm}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.5]{fig3.png}

\caption{Classification Result}

\end{center}

\end{figure}

\vspace{1 cm}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.5]{fig4.png}

\caption{Detailed Accuracy}

\end{center}

\end{figure}

\vspace{1 cm}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.8]{fig5.png}

\caption{Confusion Matrix}

\end{center}

\end{figure}

The rows show the actual instances and the column shows the predicted instances. "a" means benign and "b" means malignant. The result shows that

$425$ instances are actually benign and predicted benign.

$18$ instances are actually benign but predicted malignant.

$10$ instances are actually malignant but predicted benign.

$229$ instances are actually malignant and predicted malignant.

\vspace{1 cm}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.5]{fig6.png}

\caption{Confusion Matrix}

\end{center}

\end{figure}

\vspace{1 cm}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.7]{fig7.png}

\caption{Association Rule Mining}

\end{center}

\end{figure}

Above are the results obtained after applying the Apriori Algorithm. Fifteen rules are generated taking Size uniformity, Shape uniformity and Class. Among them rules number $2$, $3$, $4$, $13$ give significant results for determining malignancy.

\vspace{1 cm}

\begin{figure}[H]

\begin{center}

**\includegraphics**[scale=.5]{fig8.png}

\caption{Association Rule Mining}

\end{center}

\end{figure}

From Attribute Selection through Infogain Evaluator, we find that size uniformity has got the highest effect.

%------------------------------------------------

**\section{Future Work}**

Our analysis is comprises with limited dataset of size $983$(with missing value) and also from a single source. We have also applied only one machine learning tool for the entire analysis. In both the aspect we have scope of improvement where we can use more dataset from different hospitals \& research institutes and apply them on different data mining tools for more accurate analysis.

%-----------------------------------------------------

**\section{Conclusion}**

Early prediction of such deadly disease can reduce the mortality rate at a high percentage. Our aim is to provide a path directed to this prediction. By proper diagnosis and proper availability of data can save many lives. We achieve $95$\% of precision, $99.5$\% of recall and $F\_1$-Score of $99.5$\%. We stipulate that our method can help avoid clinical false negatives by performing consistency checks and provide physicians with decision support.

%----------------------------------------------------------------------------------------

% REFERENCE LIST

%----------------------------------------------------------------------------------------

\begin{thebibliography}{00}

\bibitem{Nassif:$2009$} Houssam Nassif, Ryan Woods, Elizabeth Burnside, Mehmet Ayvaci, Jude Shavik and David Page, Information Extraction for Clinical Data Mining: A Mammography CAse Study, $2009$.

\bibitem{Lyon IAfRoC}Lyon IAfRoC, World Cancer Report. International Agency for Research on Cancer Press $2003:188-193$.

\bibitem{Elattar}Elattar, Inas. “Breast Cancer: Magnitude of the Problem”,Egyptian Society of Surgical Oncology Conference, Taba,Sinai, in Egypt $(30$ March – $1$ April $2005$).

\bibitem{Daniel}Daniel F. Roses $(2005)$. Clinical Assessment of Breast Cancer and Benign Breast Disease, In: Breast Cancer: Vol. $2$, Ch. $14$, M. N. Harris [editor], Churchill Livingstone, Philadelphia.

\bibitem{Frank}Frank, A. \& Asuncion, A. $(2010)$. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California,School of Information and Computer Science.

\bibitem{Agrawal}Agrawal, R., Imielinski, T., and Swami, A. N. $1993$. Mining association rules between sets of items in large databases. In Proceedings of the $1993$ ACM SIGMOD International Conference on Management of Data, $207-216$.

\bibitem{Elsayad}A.M.Elsayad and H.A.Elsalamony.Diagnosis of Breast Cancer using Decision Tree Models and SVM.International Journal of Computer Application$(0975-8887)$,Volume $83$-No $5$,December $2013$.

\bibitem{Ahmad}Ahmad LG, EshlaghyAT, PoorebrahimiA, Ebrahimi M and RazaviAR. Using three machine learning techniques for predicting Breast cancer recurrence.J Health Med Inform $2013,4:2$

\bibitem{Shajahaan}S.S.Shajahaan, S.Shanthi, V.ManoChitra. Application of Data Mining Techniques to Model Breast Cancer Data.International Journal of Emerging Technology And Advanced Engineering,ISSN $2250-2459$,ISO $9001:2008$ Certified Journal, Volume $3$, Issue $11$, November $2013$.

\bibitem{Samar}Samar Al-Qarzaie, Sara Al –Odhaibi, Bedoor Al-Saeed and Dr.Mohammed Al-Hagery. Using the Data Mining Techniques for Breast Cancer Early Prediction, $2013$.

\bibitem{Chaurasia}V.Chaurasia and S.Pal.Data Mining Techniques:To predict and resolve breast cancer survivabilty.International Journal of Computer Science and Mobile Computing,vol.$3$ Issue $1$,January-$2014$.

\bibitem{Singhal}S.Singhal and M.Jena .A study on Weka tool for Data preprocessing ,classification and clustering.IJITEE ,Vol-$2$,May $2013$.

\bibitem{UCI}UCI machine Learning Repository http://archive.ics.uci.edu/ml/

\bibitem{WEKA}WEKA- Data Mining Software in Java.

\bibitem{Bellaachia}A.Bellaachia and ErhanGuven.Predicting Breast Cancer Survivability us-ing Data Mining Techniques.Dept.of Computer Science,The George Washington University.Washington DC.

\end{thebibliography}

%----------------------------------------------------------------------------------------

\end{multicols}

\end{document}