RESEARCH ARTICLE

Breast Cancer Statistics and Prediction Methodology: A Systematic Review and Analysis

Ashutosh Kumar Dubey*, Umesh Gupta, Sonal Jain

Abstract

Breast cancer is a menacing cancer, primarily affecting women. Continuous research is going on for detecting breast cancer in the early stage as the possibility of cure in early stages is bright. There are two main objectives of this current study, first establish statistics for breast cancer and second to find methodologies which can be helpful in the early stage detection of the breast cancer based on previous studies. The breast cancer statistics for incidence and mortality of the UK, US, India and Egypt were considered for this study. The finding of this study proved that the overall mortality rates of the UK and US have been improved because of awareness, improved medical technology and screening, but in case of India and Egypt the condition is less positive because of lack of awareness. The methodological findings of this study suggest a combined framework based on data mining and evolutionary algorithms. It provides a strong bridge in improving the classification and detection accuracy of breast cancer data.

Keywords: Breast cancer - incidence and mortality - prevalence - data mining - evolutionary algorithms

Asian Pac J Cancer Prev, **16 (10)**, 4237-4245

Introduction

However, advanced and modern world may be today and are riding on the chariot of advancement, but the truth is that there are many things which are beyond our control even today and one of them is cancer. It is wrong to say that the cancer is incurable. But the death rate of patients of breast cancer is still very high. World Health Organization (WHO) has stated that the breast cancer is most frequently found cancer in the women and it is adversary affecting millions of women all over the world. But the positive trend is that the death rate is gradually declining after 1990 due to screening, early detection, awareness and continuous improvement in treatment (Breast Cancer Deadline 2020). Some of the key risk factors of breast cancers are age, gender, affluence, family history, breast conditions, alcohol consumption and obese(Breast Cancer Deadline 2020; Report to the nation-breast cancer).

It is estimated that 1.67 new cancer cases came to light in 2012 (25% of all cancers) and 1.38 million new cancer cases detected in 2008 (23% of all cancers) according to GLOBOCAN 2012 report. Breast Cancer is very common in the region of developing and developed countries. It ranks second after the lung cancer in more developed region (15.4% of all cancers) and it is most frequent death cause in the less developed region (14.3% of all cancers) (Shah, 2014). Approximately 883,000 and 794,000 new cases were estimated in less developed and more developed regions respectively in 2012. It shows the cases of breast cancer are more as compared to 2008 (690,000 cases) (Jemal et al., 2011; Ferlay et al. 2014; Al-Darwishet al., 2014). The rate of incidence change from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women in Western Europe, 27 per 100,000 in Middle Africa and Eastern Asia to 96 in Western Europe and are high (greater than 80 per 100,000) in developed regions of the world (except Japan) and low (less than 40 per 100,000) in most of the developing regions (Jemal et al., 2011; Zhu et al., 2011; Ferlay et al., 2014).

Above mentioned statistics clearly show the horrible situation of breast cancer. So the first objective of this paper is to make survey and analysis of breast cancer incidence and mortality of different countries to find out the survival and death rates. The second objective is to bring to light the advantages and disadvantages of those algorithms and methods which can help in predicting breast cancer at the initial stage. This step will be helpful in the development of a new framework which will prove to be a milestone for cancer detection at early stage. As in case breast cancer is detected in early stage the chances of curing will be bright.

Materials and Methods

To understand the impact of breast cancer four countries named the UK, US, India and Egypt have been taken for the study from different continent.

The following sources are used for data collection of

[1]Department of Computer Science & Engineering, JK Lakshmipat University, Jaipur, India *For correspondence: ashutoshdubey123@gmail.com

incidence and mortality rates from different countries: National Cancer Intelligence Network (NCIN), UK: NCIN coordinate to develop analysis for improving clinical outcomes, cancer care and prevention. It is a part of Public Health England from April 2013.

Office for National Statistics (ONS), UK: ONS is a recognized statistical institute of the UK. It collects and publishes statistics related to population, society and economics.

Northern Ireland Cancer Registry (N. Ireland Cancer Registry): It was established in 1994, which is located in Centre for Public Health, Queen's University Belfast. It register and maintain cancer incidence and mortality. This agency is funded by the Public Health Agency for Northern Ireland.

Welsh Cancer Intelligence and Surveillance Unit (WCISU), Wales: WCISU is a national cancer registry of the Wales. It stores and publishes cancer incidence record in Wales.

Information Services Division (ISD), Scotland: ISD is a part of the NHS National Services Scotland. It conveys and gives data (Health) which are accessible for all without any fee.

Surveillance, Epidemiology, and End Results (SEER), US: The SEER is an authoritative source of information on cancer incidence and survival in the United States.

Communicable Disease Centre (CDC), US: CDC is a health promotion and prevention agency. The target of CDC is to find the purposes behind death with respect to developments, upgrading well-being methodologies and sponsorships for the perception and the investigation of study of disease surveillance and epidemiology.

World Health Organization (WHO), Switzerland: It was established in 1948 in Genera, Switzerland. WHO is an organization of the United Nation to manage and regularize well-being matters.

International Agency for Research on Cancer (IARC): IARC is a specialized cancer research agency of WHO. The fundamental point of this office is to enhance the aversion measures, distinguish malignancy at the most punctual stage and publish cancer incidences from time to time.

GLOBOCAN 2012: GLOBOCAN 2012 is a project of IARC and WHO. The main aim of this project is to estimate incidence, mortality and prevalence at the National level for 184 countries.

Epidemiology is a study of causes of any particular disease which spread out on large scale among people. So far, as cancer epidemiology is concerned, it is the study of causes and investigation of the population exposing to its risk. It means that attention is not centered on a single patient, but it is focused on the whole community to identify cancer causes and its effects in the future. The epidemiology results are not based on random assumptions; instead these are based on experiments, predictable patterns and observations, etc. The main objective of cancer epidemiology is to find the subgroups from the population, which is at the higher risk of cancer. It means epidemiology helps in the following points: *i*) To identify health problems concerning cancer. *ii*) To detect the growth of disease in the community. *iii*) To get knowledge about the risk factors of cancer or other disease. *iv*) Effect of cancer or other disease on health.

There are some indicators of epidemiology which reflect the process and outcomes on the basis of survey and existing study. It is a systematic procedure or method. Quantitative indicators or methods can be useful as these are purely based on calculation and numerical information. Some of the useful quantitative indicators are incidence, prevalence and mortality rates, etc. Incidence and mortality rates are used in this study to highlight the number of cases along with the death rates and better understood the impact of breast cancer on the country's population. The incidence rate is the calculation of finding new cases of breast cancer according to the current year population at the risk (Rothman et al., 2008; Nelson et al., 2012). The Mortality Rate (MR) is the estimate of death because of the breast cancer in the total population.

Incidence Rate (IR) can be calculated by the below formula (Rothman et al., 2008; Nelson et al., 2012): IR= (BCF/PRC) *10^n; BCF=Number of new breast cancer cases found in the current period; PRC= Total population at the risk in the current period; N=1, 2, 3….n [per 100, per 1000 population ranking]

Mortality Rate (MR) can be calculated by the below formula (Rothman et al., 2008; Nelson et al., 2012): MR= (DF/TP) * 10^n; DF=Number of death cases found in the current period; TP= Total population in the current period N=1, 2, 3….n [per 100, per 1000 population ranking] The current period means the population over the course of one year.

Results

The female breast cancer incidence and mortality rates from 1975-2014 of the UK are considered first. Age is an imperative number in case of breast cancer (Lawrence et al., 2011). So the data considered here is strongly based on the age standardized rates, so no variations are reflected in the age structure. The incidence and mortality rates Females, UK 1975-2014 are shown in Figure 1. It was calculated based on new cancer cases found in the European age-standardised rates per 100,000 populations in the UK. The incidence rates were increased in the period of 1975-2010. It is evidently clear from Figure 1 that the rates of mortality were gone high in 1975-1985, but it is gone down from 1990-2014 because of several reasons like better screening, therapies and medical care (Kingsmore et al., 2003; Autier et al., 2010). The female death rate(2010-2012) in the UK at the age of 50-69 is 34 % [National Cancer Intelligence Network UK, 2009; Office for National Statistics, 2012]. It means the risk is increased at the higher age in comparison to the younger age.

The female breast cancer incidence and mortality rates from 1975-2014 of US are considered next. The rates of incidence and mortality are shown in Figure 2. It is calculated based on new cancer cases found in the age-adjusted rates per 100,000 Populations, in US in year 1975-2014 Surveillance, Epidemiology, and End Results (SEER) Incidence. The incidence rates were increased in the period of 1985-2000. The mortality rates were increased in 1975-1990. But it is decreased from 1995-2010. Overall survival in the case of breast cancer is good in US. This is because of improved medical technology, awareness and early detection through screening. But instead of this fact the third leading cause of cancer death in the US is because of breast cancer (SEER*Stat Database, 2014).

The female breast cancer incidence and mortality rates from 1980-2014 of India are considered next. There is non- availability of incidence and mortality data in case of India. Based on the GLOBOCAN project report the incidence and mortality rates were shown in table 1. It is calculated based on new cancer cases found in the age-adjusted rates per 100,000 Populations in India. It is found that the population of India is large however the incidence rate here is less compare to other countries (Fertay et al., 2004). Lack of screening program and lack of awareness are responsible for this (Fertay et al., 2004). The death per incident ratio is highest in India, at almost 50%, compared to 30% in China and 18% in the US (Fertay et al., 2004). In terms of developed and developing countries earlier detection rate is low in India. Breast cancer death rate is high in India even with the good treatments (Fertay et al., 2004).

According to National Cancer Institute: Cairo, breast cancer is the third leading cancer (13.15 %) in Egypt (Elatar, 2002). There are very less percentage (13.3%) of breast tumor patient had mammography executed as an analytic technique (Salem et al., 2010. In case of Egypt the incidence and mortality rates of 2012 are only considered because of non-availability of data. In the above stated year 18,660 new cases were estimated there and out of this the mortality rates was estimated at 7,161. This indicates the fact that the situation of breast cancer in Egypt is almost worst like India.

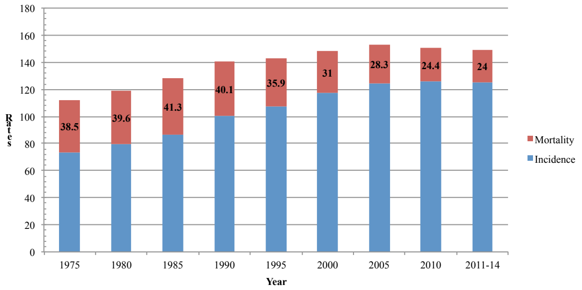

Figure 1. Breast Cancer, European Age-Standardized Incidence and Mortality Rates per 100,000 Populations, Females, UK 1975-2014. (National Cancer Intelligence Network UK, 2009; Office for National Statistics, 2012; Northern Ireland Cancer Registry, 2012; Welsh Cancer Intelligence and Surveillance Unit, 2012; Information Services Division (ISD) Scotland, 2012; Westlake et al., 2008)

Table 1. Breast Cancer, All Ages, Incidence and Mortality Rates per 100,000 Populations, Females, India 1980-2014 (Ferlay et al. 2014)

Year	Incidence	Mortality
1980	110	_
1985	115	_
1990	108	_
1995	120	_
2000	102	_
2005	108	_
2010	115	53
2011-2014	145	70

How many females are approximately being died can be calculated for the comparison between UK, US, India and Egypt. It is calculated by Incidence/(Mortality). The incidence and mortality rates pertaining to the period of 2011-2014 have been considered for death rates comparison as shown in table 2.

Due to advanced technology and awareness programs the mortality rate has decreased in the US and the UK. It is evidently clear from table 2 that the death rate of breast cancer is highest in India. It seems from this that the future move is not very easy for India to eradicate breast cancer. In case the effective steps to uproot breast cancer are taken now, even then it will take 20-30 years or more to end this situation in India.

The survival situation can be understood also by prevalence. Prevalence of cancer is such a measurement as it shows the likelihood of this disease in the population. Prevalence of breast cancer tells how many people are surviving of breast cancer disease in the current population. It can be calculated by the below formula (Coldman et al.,1992): ***Prevalence= (Total no. of cases of breast cancer)/Population***

Total no. of cases of breast cancer = newly diagnosed breast cancer + already living with breast cancer

There is a strong relationship between incidence and prevalence which depends on the history of the cancer. When incidence rate increases the prevalence rate also generally go up as it is clear from Figure 3 which is based on GLOBOCAN 2012 report. But it cannot be taken for

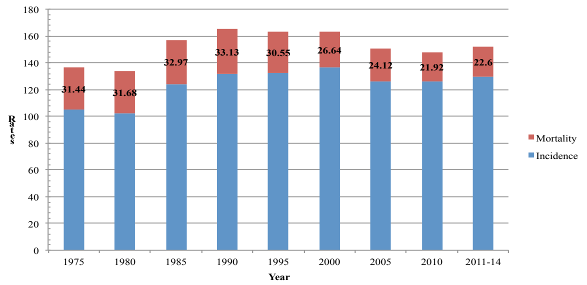

Figure 2. Breast Cancer, All Ages, All Races, SEER Incidence and Mortality Rates per 100,000 Populations, Females, US 1975-2014. (US Cancer Statistics Working Group, 2010; SEER*Stat Database, 2014; Ferlay et al., 2010; Edwards, 2014)

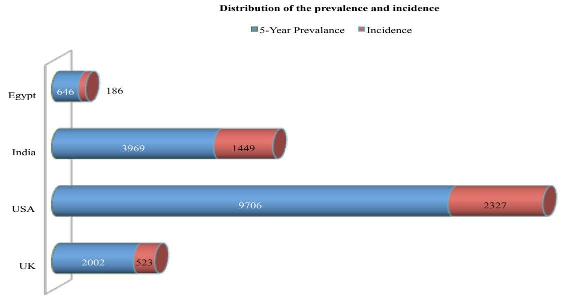

Figure 3. Distribution of the Prevalence and Incidence of Breast Cancer

Table 2. Death Rate Comparisons

Country	Incideance Mortality	Explanation
UK	5.22 (5)	In UK, for every 5 females newly identify with breast cancer, one female is dying of it.
US	5.73 (6)	In US, for every 6 females newly identify with breast cancer, one female is dying of it.
India	2.07 (2)	In India, for every 2 females newly identify with breast cancer, one female is dying of it.
Egypt	2.60 (3)	In Egypt, for every 3 females newly identify with breast cancer, one female is dying of it.

Table 3. Summary of Studies

First Author reference and Publication Year	Methods	Results
Chou et al., 2004	Artificial neural networks and multivariate adaptive regression splines	Classification accuracy rate was 97.2 %.
Tewolde et al., 2007	PSO	Classification accuracy rate was 97.4 %.
Wang et al., 2009	Structured SVM	Classification accuracy rate was 97.0 %.
Pang et. al., 2010	Impact Analysis-A Statistical Approach	The Breast Cancer Incidence rate was increased from the age of 15.
Karnan et al., 2010	Roughset, GA and ACO	Detection accuracy rate was 0.948 for 322 images.
Modiri et al., 2011	PSO	Average error rate was smaller than 5%.
Sbeity et al., 2011	Hybrid optimization model	Success parameters of optimization are objective function and the control parameters
Liu et al.,2011	Discrete Particle Swarm Optimization (DPSO) and Rule Pruning	Classification accuracy rate was 97.23 %.
Einipour et al. ,2011	Fuzzy and ACO	Classification accuracy rate was 96.99 %.
Wang et al., 2012	Association Rule based SVM Classifier	Average Classification accuracy rate was 88.83 %.
Zibakhsh et al., 2013	Memetic algorithm with a multi view fitness fuzzy approach	Classification accuracy rate was 69.43 %.
Li et al., 2013	ACO based dimension reduction method	The number of selected gene was 85 %.
Yu et al., 2013	ACOSampling with SVM classifier	Performance was approximately 94.18 for the colon dataset.
Shrivastava et al.,2013	WEKA Tool	Classification accuracy rate was 98 %.
Ahmad et al. ,2014	Random forest classifier	Classification accuracy rate was 72 %.
Martínez-Ballesteros et al., 2014	Multi-Objective evolutionary quantitative association rules	Their precision result was 100 %.
Khan et al., 2014	AKaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Deviance Information Criteria (DIC)	The standard deviation was 84.14±75. 25 months for male.
Shen et al.,2014	Feature selection method and SVM	Classification accuracy rate was 92 %.

granted. In case cure rate is good the prevalence rate will also automatically be good. In case cure rate is low then it will also affect prevalence rate adversely. In case cure rate is good but survival rate is not so the prevalence will go down. According to the report of GLOBOCAN 2012 the five year prevalence of the above stated countries are shown in Figure 3. It indicates that the highest increase of prevalence in case of USA and lowest in case of India. It proofs the same thing which is already been discussed based on table 2.

Discussion

The basic principles and measurement can be better understood by epidemiology, but due to limited resources, small area of investigation and small sample size, etc., steps will be taken towards computational methodology. Computational methodology can be helpful in data determination, detection probability, sample size expansion and classification, etc. So that in this section survey is being made on research work made earlier to detect cancer at the early stage. Pros and cons have also been discussed in this survey.

Pattern finding plays the most important role in any type of cancer diagnosis, so hidden pattern discovery is being discussed now. There is an important role of data mining in knowledge discovery. Data Mining is the process of extracting useful information from the huge database (Agrawal et al., 1994). On the basis of data type, data mining can be mainly divided into two models:

i) Descriptive: It is used to deal with the general properties. Some important descriptive functions are as under.

Association Rule Mining (ARM): It is a process of finding relationship among data (Kamsu-Foguem et al., 2013).

Clustering: To form group of similar objects. Sequence Mining: To find the relevant pattern from data in sequence. String Mining is a good example in DNA sequences.

ii) Predictive: The main purpose is to predict the class label. Models which are derived from it depend on the training data set.

Classification: To predict the class object on the basis of derived model can be formed and classes can be differentiated. Regression: The dependent and independent variables relationship can be predicted by regression.

Prediction: To automatically predict the classes based on training set.

The methodologies of breast cancer detection that is based on descriptive and predictive techniques have been discussed now. First, the methodologies have been discussed and analysed which are used the above techniques separately.

The above two models are nearly used in every areas of healthcare databases for knowledge discovery. Data mining tasks like association rule mining, correlation, sequence classifier and clustering may provide a real solution to discover similar types of group, group pattern, and frequency of the items present in the group, extraction of the significant pattern and the pattern visualization. (Khaing, 2011). Association rules mining is appropriate for finding factors which contribute to heart diseases in males and females (Nahar et al., 2013), may also be used in the case of finding breast cancer factors. It is also suggested that the ARM addresses the issue of over fitting the training data by removing the irrelevant terms from the rule, and improves the predictive power of the rule, and in the meantime simplifies it (Witten et al., 2005; Wang et al., 2007). The conventional pruning procedure is taken out at a time to examine the rule quality (Karnan et al., 2010), for the rule that there are multiple limitation conditions in one attribute, the influence of the individual parameter inside each attribute is overlooked, and thus it is worth examining each parameter separately (Witten et al., 2005; Wang et al., 2007; Liu et al., 2011). Pang et al. (2010) examined how the age, year and sex affect the probability of getting breast cancer. For this they proposed impact analysis approach. Impact analysis proceeds towards the association of the impact factor to generate calculated probability. They used Australian Institute of Health and Welfare between 1983-2003 as the dataset for breast cancer incidence and mortality. They suggested three models. The first model interprets effect of sex. Second, model interprets the probability of getting the risk of breast cancer in different age considering male subset data. Third model interprets the probability of getting the risk of breast cancer in different age considering female subset data. Their outcomes demonstrated that frequency rate of breast disease in ladies go high from the age of 15. Malpani et al. (2011) developed two different association rule mining approaches for discovering breast cancer regulatory mechanisms.

Breast cancer regulatory mechanisms of gene module, position value and position weight matrix was considered the key factor. Their data pre-processing task involved with two independent data sources: 1) a single breast cancer patient profile data file, 2) a candidate enhance information data file. Data pre-processing with two different association rule mining approach are used to identify the association rules which justify the threshold value. The threshold value is used for pruning the data and use in the experiments. It discovers the breast cancer regulatory mechanisms of gene module. The symptoms of breast cancer are not the same in all patients, it is very necessary to characterize them and give separate treatment. So clustering or classification techniques may be used. Classification is also needed as that there are several factors like age, sex, gender; hereditary, alcohol intake and overweight can affect breast cancer. The classification rule mining and decision rules are discovered through training data (Liu et al., 2011). Means it can train the cluster data according to the risk factors for better classifications. Bennett et al. (1992) proposed neural networks based linear programming for classification breast cancer data. They used Wisconsin Breast Cancer dataset (Blake et al., 2007) and reported 98.3% classification accuracy. Multi-surface outline with direct programming strategy can likewise attain to better characterization precision (Wolberg et al., 1990). According to Kerhet et al. (2006) Breast cancer is the common cause of death among women because of malignant tumors, whereas early detection leads to longest survival or even full recovery. They proposed a 3-D method, whose output is transformed to a posteriori probability of tumour presence based on Support Vector Machine (SVM) classifier. This approach was considered for noisy environments. Their results comes about around the tumor as indicated by the same structure demonstrate that it emerges against the foundation of general likelihood values. Shrivastava et al. (2013) evaluated classification method on breast cancer data. They used WEKA tool. It is a great collection of machine learning sets which are benevolent in the data mining task and undertakings. Based on the WEKA tool the classification accuracy achieved was 98 %. Ahmad et al. (2014) classified breast cancer lesions using random forest. The breast cancer lesions are obtained from fine needle aspiration (FNA). They considered approximately 700 data. It consists of 458 benign cases and 241 malignant cases. They achieved 72 % accuracy by using random forest classifier. A legitmate fitness function is indispensable which is applied for precise classification and detection (Machraoui et al., 2013). It is tough because of the data set nature. For example if the detection of the cancerous dataset from images, then improved edge detection technique is required. On the off chance that it is an instance of location of harmful tumors in mammograms, then it is exceptionally troublesome in light of the fact that carcinogenic tumors are consolidated in typical breast tissue structures (Dheeba et al., 2011). If it is the case of finding the disease changes in the data that is raw genotype (Yang et al., 2013). If it is the case of complex problems where decision variables are in large numbers (Pandey et al., 2013). If it is the case of malignant tissue finding from normal and cancer associated tissues (Zakharchenko et al., 2011).

Because of the changes in the classification and detection type, only one method is not sufficient and successful. So there is a need of the hybrid framework that is the combination of optimized classifier and prediction system. Mostly good results have been found by the hybridization of several methodologies because the symptoms of breast cancer can be of different types and these can be treated separately. So the combined methodologies can provide better results. Now methodologies discussed above with the Evolutionary Algorithms have been discussed and analysed. Evolutionary algorithms are used to find the nearer or optimal solution even in the case of complex problems. According to Wang et al. (2009) SVM is a good option for breast cancer data. They suggested a structured SVM model to determine mammographic region for normal or cancerous by accounting the group (cluster) structures in the training set. For experimentation Digital Database for Screening Mammography (DDSM), is used. This deterrent extracts yoke pan bearing texture, Gabor, curvilinear, and multi-resolution features. Their results showed better detection performance in area under the curve. SVM is also helpful in breast cancer diagnosis. The classification property of the SVM was capable of modification which turn away of the neural worrisome, and the SVM has a quite shorter training time (Chang et al., 2003). Established SVM hither respecting self-esteem resident proficiency in disorder diagnosis and result in good classification accuracy (Akay, 2009). To detect the boundary of the breast pattern Fuzzy sets, pulse coupled neural networks (PCNNs), and support vector machine, in conjunction with wavelet-based feature extraction enhance the contrast of the input images(Hassanien et al., 2012). For developing an Intelligent breast cancer classifier, multilayer perceptron (MLP) using back-propagation algorithm, probabilistic neural networks (PNN), learning vector quantization (LVQ) and support vector machine (SVM) can be a better option(George et al., 2012).

Their results show that the predictive ability of probabilistic neural network and support vector machine was stronger than the learning vector quantization and multilayer perceptron. Wang et al. (2012) proposed a novel cancer selection process based on Association Rule and SVM classifier (AR-SVM). Their classifier combines support vector machine and association rule. This method achieves high classification accuracy and good biological interpretability. The experimental results by AR-SVM show better classification accuracy in comparison to the Refined Classification Based on TopkRGS (RCBT). Some standard methods for cancer classification and prediction are Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machine (SVM), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) (Sousa et al., 2004).

Discrete Particle Swarm Optimization (DPSO) is an alternate of PSO. According to Yeh et al. (2009) DPSO can show better breast cancer classification accuracy. Liu et al. (2011) evaluated (DPSO) with rule pruning approach for finding breast cancer. They used Wisconsin breast cancer data from the University of California (UCI), Irvine machine learning repository. Their accuracy results proved that the DPSO (97.23 %) is better than PSO (97.06). Shen et al. (2014) intended to form a prediction model of breast cancer by using feature selection method and the support vector machine. They accuracy with this feature selection method was 92 %. It shows the consistency contribution of the raked features. Tewolde et al. (2007) proposed Particle Swarm Optimization (PSO) based method. It is used for single and multi-surface based data separation for breast cancer data classification. It exploits the ingenuousness, efficaciousness and flexibility of the PSO method. It subdued the PSO-based classifiers according to pre-defined statistics rupture methods, necessity regard of the dataset for offing. Their classification accuracy results are in the range of 97-100 %. The frequency range and measurement setup highly affects the electrical properties (Cho et al., 2006; Lazebnik et al., 2007). The planar brandish recklessness work breech be modelled by prearrange radiators, which are the most common for the radiometry and multi-structure (Orfanidis et al., 2002; Thakur et al., 2002). Modiri et al. (2011) used PSO algorithm to estimate the sustainibility of the kinfolk layers as in (Yeung et al., 2009; Gandhi et al., 2010) at microwave frequency band. They have achieved improved final responses accuracy.

The integrated performance of neural networks with the multivariate adaptive regression splines (MARS) was studied (Chou et al., 2004). The results show good classification accuracy. Finite-difference frequency domain and particle swarm optimization was proposed to reconstruct the breast cancer cell dimension and determines its position (Zainud-Deen et al., 2008). It uses 2-D and 3-D models. The results prove the possibility for attaining the localization and the size of the cancer inside the breast. Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Deviance Information Criteria (DIC) were adopted to measure the goodness of fit for Modelling Male Breast Cancer (Khan et al., 2014). The sample database of 500 males has been selected for this modelling. The results suggest the exponential Weibull model fits the male survival data (Khan et al., 2014).

Martínez-Ballesteros et al. (2014) proposed a multi-objective evolutionary algorithm for mining association rules. It is the process to discover more relevant genes. The name of this methodology is Gene Association Rule Network (GarNet). It is applied on yeast cell cycle. There are two steps to deal with the generation of the gene association networks. First, GarNet carries out an inference method based on an Iterative Rule Learning (IRL) to extract gene–gene associations. It develops the network through Quantitative Association Rules (QAR) found in several input microarray datasets and found the consistent results. The effectuation study of this fact finding was performed as like in the study of Gallo et al.(2011). For predicting gene function, DNA microarray is a good choice (Zhou et al., 2002). DNA microarray is also an obvious choice for Gene regulatory mechanism (Husmeier, 2003; Segal et al., 2003), drug discovery (Evans et al., 2004) cancer classification (Alon et al., 1999; Conrads et al., 2003) mining specific tumours (Golub et al., 1999; Wigle et al., 2002; Nutt et al., 2003). There are several gene selection approaches were developed including Statistical Analysis (Khan et al., 2001), Bayesian Mode (Lee et al., 2003), Fisher Linear Discriminator Analysis (Xiong et al., 2001) and support vector machine (Guyon et al., 2002; Tang et al., 2006). Zibakhsh et al. (2013) evaluated DNA microarray-based gene expression profiles to find cancerous and benign tumours from the gene. According to Zibakhsh et al. (2013) classifiers like K-Means, SVM and ANN experience low comprehensibility. So they suggested memetic algorithm based method. Memetic algorithm is capable of finding fuzzy rules from cancer data. Its fitness function considers two different evaluating procedures. The first procedure is located in the main evolutionary structure of the algorithm. It evaluates each single fuzzy rule. The second procedure evaluates the quality of each fuzzy rules from the whole set. So the above method enhances the discovery process by evaluating each fuzzy rule. Karnan et al. (2010) suggested microcalcifications as the key symptoms facilitating early detection of breast cancer. They suggested textural features. t is suggested stranger the articulated mammogram image. It was hand-me-down to bracket the microcalcifications. The classification classifies it into benign, malignant or normal. They used rough set based reduction algorithms such as Heuristic approach, Hu's algorithm, Quick Reduct (QR), and Variable Precision Rough Set (VPRS) and the metaheuristic algorithms such as GA and HPACO algorithms (Goldberg, 1989; Kim et al., 1999; Hassanien et al., 2004). Right the textural Phiz were normalized between zero and one. The normalized imperturbability were old as an input to a three-layer Back Propagation Network (BPN) to classify the microcalcifications into benign, malignant or normal. to set the microcalcifications into congenial, hateful or accustomed. The garner logic were go by trained network. It is 0.9 for malignant, 0.5 for benign and 0.1 for normal images. Receiver Operating Characteristics (ROC) was used as the classification performance evaluator. The competence was tested on 161 pairs of digitized mammograms and proof that ACO outperforms. Sbeity et al. (2011) suggested mathematical model for the optimal solution finding in the direction of cancer therapy. The model was considered for an individual patient over a fixed time horizon. According to Sbeity et al. (2011) the success of any optimization function depends on the optimization procedure and the control parameters. Their results are more accurate in optimal treatment. Einipour et al. (2011) proposed Fuzzy and ACO based method for breast cancer detection. The fuzzy rules are used for the interpretability capability and ACO optimized the fuzzy rule sets. The datasets was taken from UCI machine learning repository. They achieved 96.99 % accuracy by using Fuzzy-ACO. Li et al. (2013) proposed a bionic optimization algorithm based dimension reduction method named Ant Colony Optimization-Selection (ACO-S) for high-dimensional datasets. The performance has been checked on the dimensions varying from 7129 to 12000. The results show that ACO-S has a notable ability to generate a gene subset with the smallest size and shows high classification accuracy. Yu et al. (2013) suggested Ant Colony Optimization (ACO) Sampling to address the problem of class imbalance problem. This algorithm starts with feature selection to eliminate noisy genes in data. The original training set is then divided into groups: training set and validation set. In each division, they modified the parameter and concluded that the ACO is efficient in finding optimal training sample subset. The importance of the corresponding majority sample can be shown by frequency list. Their approach outperforms in generating stronger generalization ability.

It has been observed by this study that the good results can be found in case of combined methodologies (Data Mining and Evolutionary Algorithms). Based on the overall study it is observed that data mining techniques have the capability to give better results in terms of classifying breast cancer dataset. It is also observed that the evolutionary algorithm can provide optimal solution. The possibility of good results also in case if nearer solution/optimal solution can be found instead of exact solution. So if data mining techniques and evolutionary algorithms are combined applied, then the chances of detection of breast cancer can be improved. The study results of table 3 also support this discussion.

Conclusions

Breast cancer has created a terrible situation in almost all over the world according to this study and discussion. It has been observed that the death rate is gradually coming down in some developed countries like the UK and US because of the developed technologies used in diagnosis and awareness. But in developing countries like India the situation is not good and some effective steps should be taken in this direction without any delay.

This study has been made on methodologies by which the breast cancer can be detected at early stages by using the breast cancer data set. It is clear from this study that the Association Rule Mining, Classification, Clustering and Evolutionary Algorithms are good at detection and classification of breast cancer data. It is also observed that if the properties of the symptoms are identified correctly, the chances of accurate detection will improve. It is also observed by the results of the previous methods that the classification algorithm increases the possibility of improved detection accuracy. The characteristics of breast cancer symptoms are different, so the chances of good results by using single algorithm are less. But by the use of combined algorithms at different levels will produce good results. So it is concluded that the framework based on data mining and evolutionary algorithms can be a milestone in case of breast cancer detection.

References

Agrawal R, Srikant R (1994). Fast algorithms for mining association rules. *VLDB*, **1215**, 487-99.

Ahmad F, Yusoff N (2013). Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. 13th International Conference on in Intelligent Systems Design and Applications (ISDA), **IEEE**, 121-5.

Akay MF (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems Applications*, **36**, 3240-7.

Al-Darwish AA, Al-Naim AF, Al-Mulhim KS, et al (2014). Knowledge about cervical cancer early warning signs and symptoms, risk factors and vaccination among students at a medical school in Al-Ahsa, Kingdom of Saudi Arabia. *Asian Pac J Cancer Prev*, **15**, 2529-32.

Alon U, Barkai N, Notterman DA, Gish K, et al (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*, **96**, 6745-50.

Autier P, Boniol M, LaVecchia C, et al (2010). Disparities in breast cancer mortality trends between 30 European countries: retrospective trend analysis of WHO mortality database. *BMJ*, **341**, 1-7.

Bennett KP, Mangasarian O (1992). Neural Network Training via Linear Programming. Advance in Optimization and Parallel Computing, 56-67.

Breast Cancer Deadline 2020 (2012). Retrieved May 10, 2014, from www.breastcancerdeadline2020.org/

Blake CL, Merz CJ (2007). UCI machine learning repository of machine learning databases. www.ics.uci.edu/~ mlearn/MLSummary.html.

Report to the nation-breast cancer (2012). Cancer Australia, Surry Hills, NSW.

Chang RF, Wu WJ, Moon WK, et al (2003). Support vector machines for diagnosis of breast tumors on US images. *Acad Radiology*, **10**, 189-97.

Cho J, Yoon J, Cho S, et al (2006). In vivo measurements of the dielectric properties of breast carcinoma xenografted on nude mice. *Int J Cancer*, **119**, 593-8.

Chou SM, Lee TS, Shao YE,et al (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, **27**, 133-42.

Coldman AJ, McBride ML, Braun T (1992). Calculating the prevalence of cancer. Statistics in medicine, 11, 1579-1589.

Conrads TP, Zhou M, Petricoin EF, et al (2003). Cancer diagnosis using proteomic patterns, *Expert Rev Mol Diagn*, **3**, 411-20.

Dheeba J, Selvi ST (2011). A CAD system for breast cancer diagnosis using modified genetic algorithm optimized artificial neural network. In Swarm, Evolutionary, and Memetic Computing, 349-57.

Edwards BK, Noone AM, Mariotto AB, et al (2014). Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer*, **120**, 1290-314.

Elatar I (2002). Cancer registration, NCI Egypt 2001. Cairo, Egypt, National Cancer Institute, Available from: http://www.nci.edu.eg/Journal/nci2001%20.pdf.

Einipour A (2011). A fuzzy-ACO method for detect breast cancer. *Global J Health Science*, **3**, 195-9.

Evans WE, Guy RK (2004). Gene expression as a drug discovery tool. *Nature Genetics*, **36**, 214-5.

Ferlay J, Soerjomataram I, Ervik M, et al (2014). GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer; 2013. Visit: http://globocan.iarc.fr.

Ferlay J, Shin HR, Bray F, et al (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*, **127**, 2893-917.

Fertay J, Bray F (2004). GLOBOCAN 2002Cancer Incidence, Mortality and Prevalence Worldwide, IARC CancerBase No. 5, Version 2.0; IARC Press, Lyon.

Gallo C. A, Carballido JA, Ponzoni I (2011). Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinformatics*, **12**, 1-21.

Gandhi KR, Karnan M, Kannan S (2010). Classification rule construction using particle swarm optimization algorithm for breast cancer data sets. In Signal Acquisition and Processing (ICSAP), 233-7.

George YM, Bagoury BM, Zayed HH, et al (2012). Breast fine needle tumor classification using neural networks. *Int J Computer Sci Issues*, **9**, 247-56.

Goldberg DE (1989). Genetic algorithms in search, optimization, and machine learning. Reading Menlo Park: Addison-wesley.

Golub TR, Slonim DK, Tamayo P, et al (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-7.

Guyon I, Weston J, Barnhill S, et al (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, **46**, 389-422.

Hassanien AE, Ali JM (2004).Enhanced rough sets rule reduction algorithm for classification digital mammography. *J Intelligent Systems*, **13**, 151-71.

Hassanien AE, Kim TH (2012). Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks. *J Applied Logic*, **10**, 277-84.

Husmeier D (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271-82.

Information Services Division (ISD) Scotland, 2012 (2012). Available from: http://www.isdscotland.org/Health-Topics/Cancer/Publications/index.asp#605

Jemal A, Bray F, Center MM, et al (2011). Global cancer statistics. *CA: A Cancer J Clin*, **61**, 69-90.

Kamsu-Foguem B, Rigal F, Mauget F (2013). Mining association rules for the quality improvement of the production process. *Expert Systems with Applications*, **40**, 1034-45.

Karnan M, Gandhi KR (2010). Diagnose breast cancer through mammograms, using image processing techniques and optimization techniques. In Computational Intelligence and Computing Research (ICCIC), 1-4.

Kerhet A, Raffetto M, Boni A, et al (2006). A SVM-based approach to microwave breast cancer detection. *Engineering Applications Artificial Intelligence*, **19**, 807-18.

Khaing HW (2011). Data mining based fragmentation and prediction of medical data. *In Computer Research and Development (ICCRD)*, **2**, 480-5.

Khan HM, Saxena A, Rana S, et al (2014). Bayesian method for modeling male breast cancer survival data. *Asian Pac J Cancer Prev*, **15**, 663-9.

Khan J, Wei JS, Ringner M, Saal LH, et al (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, **7**, 673-9.

Kim JK, Park H (1999). Statistical textural features for detection of microcalcifications in digitized mammograms. *IEEE Transactions on Medical Imaging*, **18**, 231-8.

Kingsmore D, Ssemwogerere A, Hole D, Gillis C (2003). Specialisation and breast cancer survival in the screening era. *Br J Cancer*, **88**, 1708-12.

Lawrence G, Kearins O, Lagord C, et al (2011). The second all breast cancer report. national cancer intelligence network: London.

Lazebnik M, Popovic D, McCartney L, et al (2007). A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries. *Physics Medicine Biol*, **52**, 6093-115.

Lee KE, Sha N, Dougherty ER, et al (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90-7.

Li Y, Wang G, Chen H, et al (2013). An ant colony optimization based dimension reduction method for high-dimensional datasets. *J Bionic Engineering*, **10**, 231-41.

Liu Y, Chung YY (2011). Mining cancer data with discrete particle swarm optimization and rule pruning. *In IT in Medicine and Education (ITME)*, **2**, 31-4.

Machraoui AN, Cherni MA, Sayadi M (2013). Ant Colony optimization algorithm for breast cancer cells classification. In Electrical Engineering and Software Applications (ICEESA), 1-6.

Malpani R, Lu M, Zhang D, Sung WK (2011). Mining transcriptional association rules from breast cancer profile data. In Information Reuse and Integration (IRI), 154-9.

Martínez-Ballesteros M, Nepomuceno-Chamorro IA, Riquelme JC (2014). Discovering gene association networks by multi-objective evolutionary quantitative association rules. *J Computer System Sciences*, **80**, 118-36.

Modiri A, Kiasaleh K (2011). Permittivity estimation for breast cancer detection using particle swarm optimization algorithm. *In Engineering in Medicine and Biology Society (EMBC)*, 1359-62.

Nahar J, Imam T, Tickle KS, Chen YPP (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, **40**, 1086-93.

National Cancer Intelligence Network and Cancer Research UK (2009). Cancer Incidence and Survival by Major Ethnic Group, England 2002-2006.

Nelson KE, Williams CM (2012). Infectious disease epidemiology. Jones & Bartlett Publishers.

Northern Ireland Cancer Registry (2012). Available from: http://www.qub.ac.uk/research-centres/nicr/CancerData/OnlineStatistics/

Nutt CL, Mani DR, Betensky RA, et al (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer research*, **63**, 1602-7.

Office for National Statistics (2012). Available from: http://www.ons.gov.uk/ons/search/index.html?newquery=cancer+registrations

Orfanidis SJ (2002). Electromagnetic waves and antennas. Rutgers University, 227-50.

Pandey B, Garg N (2013). Swarm optimized modular neural network based diagnostic system for breast cancer diagnosis. *Int J Soft Computing, Artificial Intelligence Applications*, **2**, 11-20.

Pang KP, Ali AS (2010). Finding association of impact factor for breast cancer patient-A novel statistical approach. *In Neural Networks (IJCNN)*, 1-5.

Rothman KJ, Greenland S, Lash TL (2008). Modern epidemiology. Lippincott Williams & Wilkins.

Salem AA, Salem MAE, Abbass H (2010). Breast cancer: surgery at the south Egypt cancer institute. *Cancers*, **2**, 1771-8.

Sbeity H, Younes R, Topsu S, Mougharbel I (2011). Comparative study of the optimization theory for cancer treatment. In Biomedical Engineering and Informatics (BMEI), 2, 927-33.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166-76.

SEER*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2013 Sub (1973-2011) U.S., National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the November 2013 submission, www.seer.cancer.gov.

Shah S (2014). BreastCancerIndia.net. Retrieved January 30, 2014, from http://www.breastcancerindia.net/.

Shrivastava S, Sant A Aharwal R (2013). An overview on data mining approach on breast cancer data. *Int J Advanced Computer Research*, **3**, 256-62.

Shen R, Yang Y, Shao F (2014). Intelligent breast cancer prediction model using data mining techniques. Sixth International Conference on In Intelligent Human-Machine Systems and Cybernetics (IHMSC), **1**, 384-7.

Sousa T, Silva A, Neves A (2004). Particle swarm based data mining algorithms for classification tasks. *Parallel Computing*, **30**, 767-83.

Tang EK, Suganthan PN, Yao X (2006). Gene selection algorithms for microarray data based on least squares support vector machine. *BMC Bioinformatics*, **7**, 1-16.

Tewolde GS, Hanna DM (2007). Particle swarm optimization for classification of breast cancer data using single and multisurface methods of data separation. In Electro/Information Technology, 443-6.

Thakur KP, Holmes WS, Carter G. (2002). An inverse technique to evaluate thickness and permittivity using reflection of plane wave from inhomogeneous dielectrics. In ARFTG Conference Digest, 1-7.

US Cancer Statistics Working Group. (2010). United States cancer statistics: 1999-2006 incidence and mortality web-based report. Atlanta, GA.

Wang D, Shi L, Ann HP (2009). Automatic detection of breast cancers in mammograms using structured support vector machines. *Neurocomputing*, **72**, 3296-302.

Wang Z, Sun X, Zhang D (2007). A PSO-based classification rule mining algorithm. In Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence, 377-84.

Wang M, Su X, Liu F, Cai R (2012). A cancer classification method based on association rules. In Fuzzy Systems and Knowledge Discovery (FSKD), 1094-8.

Welsh Cancer Intelligence and Surveillance Unit (2012). Available from: http://www.wales.nhs.uk/sites3/page.cfm?orgid=242&pid=51358

Westlake S, Cooper N (2008). Cancer incidence and mortality: trends in the United Kingdom and constituent countries, 1993 to 2004. *Health Stat Q*, **38**, 33-46.

Wigle DA, Jurisica I, Radulovich N, et al (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res*, **62**, 3005-8.

Witten IH, Frank E (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wolberg WH, Mangasarian OL. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci USA*, **87**, 9193-6.

Xiong M, Li W, Zhao J, Jin L, Boerwinkle E (2001). Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics Metabolism*, **73**, 239-47.

Yang CH, Lin YD, Chuang LY, Chang HW (2013). SNP barcodes generated using particle swarm optimization to detect susceptibility to breast cancer. *Natural Science*, **5**, 359-67.

Yeh WC, Chang WW, Chung YY (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications*, **36**, 8204-11.

Yeung CW, Leung FF, Chan KY, Ling SH (2009). An integrated approach of particle swarm optimization and support vector machine for gene signature selection and cancer prediction. *IJCNN*, 3450-6.

Yu H, Ni J, Zhao J (2013). ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing*, **101**, 309-18.

Zainud-Deen SH, Hassen WM, Ali EM, et al (2008). Breast cancer detection using a hybrid finite difference frequency domain and particle swarm optimization techniques. *In Radio Science Conference*, **2008**, 1-8.

Zakharchenko O, Greenwood C, Alldridge L, Souchelnytskyi S (2011). Optimized protocol for protein extraction from the Breast Tissue that is compatible with Two-Dimensional Gel electrophoresis. *Breast cancer: basic and clinical research*, **5**, 37-42.

Zhou X, Kao MCJ, Wong WH (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci USA*, **99**, 12783-8.

Zhu YY, Zhou L, Jiao SC, Xu LZ (2011). Relationship between soy food intake and breast cancer in China. *Asian Pac J Cancer Prev*, **12**, 2837-40.

Zibakhsh A, Abadeh MS (2013). Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. *Engineering Applications of Artificial Intelligence*, **26**, 1274-81.