

A Survey on Breast Cancer Scenario and Prediction Strategy

Ashutosh Kumar Dubey, Umesh Gupta, and Sonal Jain

JK Lakshmipat University, Jaipur, India
ashutoshdubey123@gmail.com,
{umeshgupta,sonaljain}@jkl.u.edu.in

Abstract. The breast cancer is one of the most critical cancer types that are found in women. Significant number of researches has been going on this disease. Post diagnosis treatments results into side effect on the patient's health. The only way to survive from this disease is its early detection. This paper is aimed to survey the frameworks proposed by researchers to detect the breast cancer at initial stage leading to its treatment with minimum side effects. The paper also suggests improvements in the direction of development of such frameworks.

Keywords: Breast Cancer, Prediction Strategy, Optimization, Data Mining.

1 Introduction

Breast cancer is by approximately the most suitable minister to murrain surrounded by women with an estimated 1.38 and 1.67 million new cancer cases diagnosed in 2008(23% of all cancers) and 2012 respectively (25% of all cancers) and ranks second overall (10.9% of all cancers)[1][4]. It is suitable the surpass regular complaint both in well-ripened and growth acumen upon surrounding 6, 90,000 new cases estimated in each region (population ratio 1:4) [3]. Incidence rates vary from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women 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 [2] [3].The range of mortality rates is much less (approximately 6-19 per 100,000) because of the most favorable survival of breast cancer in (high-incidence) developed regions [3]. From the overall ratio breast cancer ranks as the fifth cause of death (458,000-deaths), but it is still the most frequent cause of cancer death in women in both developing (269,000-deaths, 12.7% of total) and developed regions, where the estimated 189,000-deaths is almost equal to the estimated number of deaths from lung cancer (188,000-deaths) [3]. The worldwide incidence and mortality is shown in table 1(incidence and mortality 2008 and 2012). Most of the cancer types can be curable if the cancer can be detected in the early stages [5]; early and precise detection is a critical factor in

selecting the proper and effective treatment for the disease [6]. In this paper, a survey has been done to find the optimal solution in the direction of breast cancer [7]. So in this regard survey of data mining and optimization techniques are also done which are notably in cancer identification and prediction [8].

Table 1. Breast Cancer Incidence and Mortality Worldwide [3] [4]

Estimated age-standardized rates (World) (per 100,000) Estimated numbers (thousands)	Cases 2008	Deaths 2008	Cases 2012	Deaths 2012
World	1384	458	1677	522
More developed regions	692	189	794	198
Less developed regions	691	269	883	324
WHO Africa region (AFRO)	68	37	100	49
WHO Americas region (PAHO)	320	82	408	92
WHO East Mediterranean region (EMRO)	61	31	99	42
WHO Europe region (EURO)	450	139	500	143
WHO South-East Asia region (SEARO)	203	93	240	110
WHO Western Pacific region (WPRO)	279	73	330	86
IARC membership (24 countries)	740	214	940	257
United States of America	182	40	233	44
China	169	44	187	48
India	115	53	145	70
European Union (EU-28)	332	89	367	91

The rest of the paper is arranged as follows: Section 2 introduces related work; Section 3 shows the problem domains and analysis; Section 4 shows the conclusions and future work.

2 Related Works

In 2007, Tewolde et al., [9] ingenuousness, efficaciousness and flexibility of the Particle Swarm Optimization (PSO) method to propose single and multi-surface based data separation methods for classification of Breast Cancer Data were suggested. In 2010, Pang et al., [10] how the age, year and sex affect the probability of getting breast cancer by applying a new impact analysis approach was suggested. The impact analysis is an approach that is used to interpret the association of the impact factor with the acceptance inconstant, and to represent the result in terms of estimated probability [10]. The prudence has demonstrated the female incidence rate of breast cancer has the sharp increase starting from the age 15. In 2011, Malpani et al., [11] a

method of data pre-processing and two different associations rule mining approaches for discovering breast cancer regulatory mechanisms of gene module were proposed. It is involved with two independent data sources: (a) a single breast cancer patient profile data file, (b) a candidate enhance information data file. It is useful also in the case of other data mining applications. In 2011, Modiri et al., [12] particle swarm optimization (PSO) algorithm was used to estimate the permittivity's of the tissue layers at microwave frequency band. In addition, breast cancer is an appropriate candidate of microwave radiometry (MWR) due to the breast's exclusive physiology. They evaluated several algorithms for analyzing the measurement data and solving the inverse scattering problem in microwave radiometry (MWR), and different levels of accuracy. They develop two distinct algorithms for the two considered scenarios. In the first scenario, they assume no prior knowledge of the tissue under the test, whereas, in the second scenario, prior knowledge is assumed. Yet, in the favor of these researches underestimate the loss encountered by the test data samples; the methods are not valid for body tissue case. In 2011, Sbeity et al., [13], optimal uses of different medicines are necessary for controlling the number of clinically observed tumor cells was suggested. For the optimal solution, finding in the direction of cancer therapy mathematical model was suggested and provide a comparative study. The accomplishing of the optimization methods depends on the choice of the objective function and its functional relationship to the control parameters [13]. In 2012, Wang et al., [14] a novel cancer selection method; Association Rule-based SVM-classifier (ARSVM) was proposed. It is a classifier that combines support vector machine and association rule. It can be used as feature extraction approach to catch the relation(non-linear) among different genes, and support vector machine is used to classify the transformed gene expression data [14]. It achieves both high classification accuracy and good biological interpretability. In 2011, Liu et al., [15] a classifier using discrete particle swarm optimization (DPSO) with an additional new rule pruning procedure for detecting lung cancer and breast cancer was proposed. It improves the classification accuracy, and it is effective in making cancer prediction. In 2012, Dubey et al., [16] an efficient method for knowledge discovery which is based on subset and superset approach was proposed. A frequent superset means it contains more transactions than the minimum support. It utilizes the concept that if the itemset is not frequent, but the superset may be frequent which is considering for the further data mining task. By this approach, improved association can be achieved. In 2013, Zibakhsh et al., [17] a new memetic algorithm was proposed. It is capable of extracting interpretable and accurate fuzzy if-then rules from cancer data. In comparison to classic memetic algorithms, it enhances the rule discovery process significantly. In 2013, Nahar et al., [18] sick and healthy factors which contribute to heart disease for males and females were investigated. Association rule mining is used to identify these factors and the UCI Cleveland dataset, a biological database, is considered along with the three-rule generation algorithms Apriori, Predictive Apriori and Tertius. The resting ECG being either normal or hyper and slope being flat are potentially high-risk factors for women only. For men, on the in rotation deal out, desolate an abstemious rule expressing resting ECG being hyper was shown to be a significant factor. So for women, resting the ECG status is a key distinct factor for

heart disease prediction. In 2013, Li et al., [19] a bionic optimization algorithm based dimension reduction method named Ant Colony Optimization-Selection (ACO-S) for high-dimensional datasets was presented. The match close-fisted role of rove ACO-S has a notable ability to generate a gene subset with the smallest size and salient features while yielding high-classification accuracy. In 2009, Wang et al., [20] structured Support vector machine (SVM) model for determining mammographic region for normal or cancerous by considering the cluster structures in the training set was suggested. It achieves better detection performance compared with a tested SVM classifier in terms of the area under the ROC curve. In 2013, Yu et al., [21] ACO Sampling method based on the idea of Ant Colony Optimization (ACO) to address the problem of class imbalance problem frequently occurs was proposed. The above study shows that there is the need of hybrid methodology as a single methodology is not sufficient or convincing for early stage detection.

3 Problem Domains and Analysis

In 2011, National Breast Cancer Coalition (NBCC) [22] issued a progress report on breast cancer named “Breast Cancer Deadline 2020”. According to the report, breast cancer will be totally uprooted by 2020 and in 90 % cases the death of breast cancer patient died due to spreading of breast cancer in other parts of body. All the breast cancer is not of the same type and their division can be done through biology of the tumor [22]. So the types can be diagnoses differently for better results. It means the symptoms of breast cancer are not the same in all patients, in view of this fact it is very necessary to characterize them and give separate treatment[22][23]. So there is the need of clustering and classification techniques. According to the estimation by National Cancer Institute (NCI) in the United States more than 288,000 women and 2,140 men developed invasive in 2011, and 39,520 women and 450 men died from the disease [24]. It can be increases in 2012 with over 290,000 women and 2,190 men predicted to receive a diagnosis [25]. In 2012, NBCC investigates to use computational and bioinformatics approaches to carry out a systematic analysis of existing and developing genomic, proteomic, glycaemic or immune system profiling data within the context of human breast cancer [22]. The detail incidence and mortality rate of US can be better analyzed by figure 1 and figure 2. The death rates shown in figure 2 shows the alarming condition. Age is also a major factor for the breast cancer. Breast cancer diagnosed in women in UK was reported in [26]. According to this report 50 % cases are diagnosed in the women in the age of 50-69. 13 % of the patients were diagnosed in the age of less than 50, 40 % diagnosed in the age between 50-69 and 47 % are diagnosed over age 70[23]. Means increasing age is also one of the strongest risk factor for breast cancer. It is shown in Figure 3. According to [27] if breast cancer is detected in the primary stage then the chances of survival is more. The chances of survival in the case small tumour diagnosis are more. The survival in the case of small tumours at diagnosis ($\leq 10\text{mm}$)[27] is around 98 % compared with 73 % for women with large tumours($\geq 30\text{ mm}$)[28][23].

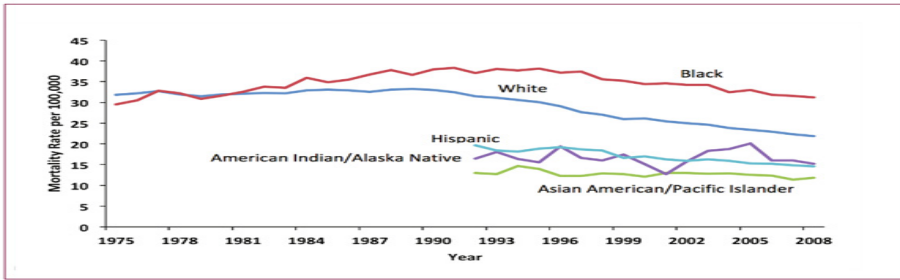


Fig. 1. Female Breast Cancer Incidence Rates by Race and Ethnicity, US, 1975-2008[22][23][25]

[Rates are age-adjusted to the 2000 US standard population]

Data source: Surveillance, Epidemiology, and End Results (SEER) Program, 1975-2008, Division of Cancer Control and Population Science, National Cancer Institute, 2012. Data for whites and blacks are from the SEER 9 registries. Data for other race/ethnicities are from the SEER13 registries. Hispanics and Non-Hispanics are not mutually exclusive from whites, blacks, Asian/Pacific Islanders, and American Indians/Alaska Natives.

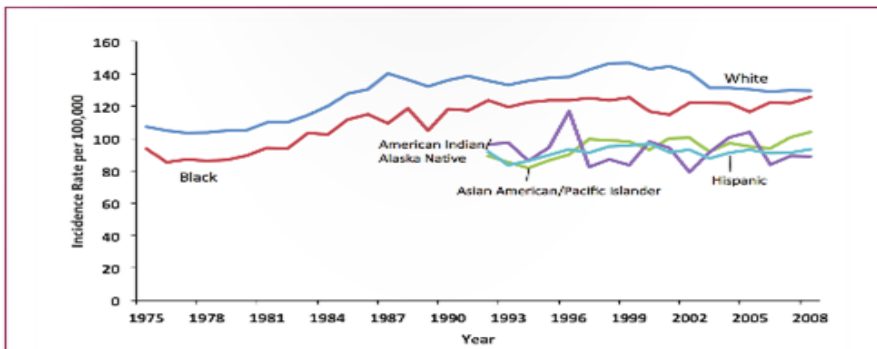


Fig. 2. Female Breast Cancer Mortality Rates by Race and Ethnicity, US, 1975-2008[22][23][25]

[Rates are age-adjusted to the 2000 US standard population]

Data source: US Mortality Files, National Center for Health Statistics, CDC. Rates for American Indian/Alaska Native are based on the CHSDA (Contract Health Service Delivery Area) counties.

Other risk factor is gender. Breast cancer chances in woman are high in comparison to the men [23]. It may affluence more frequently in western population. It had been analyzed that breast cancer cases can be high if it is a hereditary disease. Regular habits of taken alcohol intake can also be the strongest cause. Women diagnosed with invasive breast cancer have an increased risk of developing another breast cancer. Dense breast tissue on mammography is also emerging as a strong risk factor [23]. Overweight or obese increase the chances of breast cancer. Area is also a

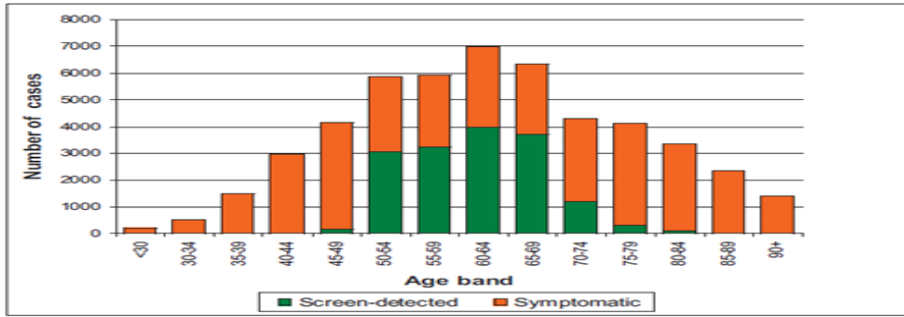


Fig. 3. Age Profile and route of presentation for women diagnosed with breast cancer in 2007[26]

major cause[23]. A survey on 30 % Australian women diagnosed with breast cancer which are lived in major cities and find the chances are high in urban area in comparison to the rural area[23].

Data mining have great potential for exploring the hidden patterns in the Text sets of the medical domain. These patterns can be utilized for clinical diagnosis. Even so, the open raw medical Texts are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user-oriented betterment. Data mining and Observations both crack opinion discovering patterns and structures in data. Statistics deals round mongrel in large quantity unequalled, whereas data mining deals with heterogeneous fields. Association rules mining which is appropriate for finding factors can be used [18]. Clustering is also helpful because of different categorization is needed and treated the category differently. Classification is also needed as there are several factors like age, sex, gender, hereditary, alcohol intake and overweight can affect breast cancer. So there is need of strong classifier. Rule pruning sermon the business 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[15][29][30]. According to[15][8][30] conventional pruning procedure taken out at a time to examine the rule quality[8], for rule which there are multiple limitation conditions in one attribute, the influence of individual parameter inside each attribute is overlooked, and thus it is worth examine each parameter separately. Optimization can be used to generate the rules. In this direction different researchers use different optimization techniques. In [15] and [8] particle swarm optimization (DPSO) was used. Genetic algorithm was suggested in [31]. It can reduce both the cost and the time of rule discovery. In [19] a bionic optimization algorithm based dimension reduction method named Ant Colony Optimization-Selection (ACO-S) was suggested. It can not only select a feature subset of smallest size, but also achieve the best classification accuracy. So as per the analysis and discussion there is the need of a hybrid algorithm which can be a combination of Association Rule Mining, Data Clustering and Optimization for early stage breast cancer detection.

4 Conclusions and Future Work

In this research paper survey of related research papers has been presented and discussed the methodology used in breast cancer detection. It is concluded that age, sex, regular habits of taken alcohol, urban/rural region and weight are the major factors of Breast Cancer. The chances of breast cancer cases can be high if it is a hereditary disease. For analysis, there is a need of huge amount of data so data mining may be used. The symptoms of breast cancer are not the same in all patients, in view of this fact it is very necessary to characterize them and give separate treatment. For grouping of alike symptoms clustering can be used. The best results in different conditions may be discovered by optimizing the results.

As per our observation and discussion there are lots of factors which affect breast cancer in different ways. So if the methodologies are clubbed in a single framework the chances of finding better solution will be increased. So a hybrid framework consists of classification, clustering, association and optimization will prove to be better in the above situation.

References

1. <http://www.breastcancerindia.net/>
2. <http://www.asianscientist.com/health-medicine/late-breast-cancer-diagnosis-linked-survival-rates-rural-india-2013/>
3. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. *A Cancer Journal for Clinicians* 61(2), 69–90 (2011)
4. GLOBOCAN: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in (2012)
5. Ott, J.J., Ullrich, A., Miller, A.B.: The importance of early symptom recognition in the context of early detection and cancer survival. *European Journal of Cancer* 45(16), 2743–2748 (2009)
6. Shital, S., Kusiak, A.: Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine* 37(2), 251–261 (2007)
7. Wei-Chang, Y., Chang, W.-W., Chung, Y.Y.: A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications* 36(4), 8204–8211 (2009)
8. Tiago, S., Silva, A., Neves, A.: Particle swarm based data mining algorithms for classification tasks. *Parallel Computing* 30(5), 767–783 (2004)
9. Tewolde, G.S., Hanna, D.M.: Particle swarm optimization for classification of breast cancer data using single and multisurface methods of data separation. In: *IEEE International Conference on Electro-Information Technology*, pp. 443–446. IEEE (2007)
10. Kwok-Pan, P., Ali, A.S.: Finding association of impact factor for breast cancer patient-A novel statistical approach. In: *Second International Conference on Environmental and Computer Science (ICECS)*, pp. 1–5. IEEE (2010)
11. Rakhi, M., Lu, M., Zhang, D., Sung, W.K.: Mining transcriptional association rules from breast cancer profile data. In: *IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 154–159. IEEE (2011)

12. Arezoo, M., Kiasaleh, K.: Permittivity estimation for breast cancer detection using particle swarm optimization algorithm. In: Annual International Conference of the IEEE EMBS, pp. 1359–1362. IEEE (2011)
13. Hode, S., Younes, R., Topsu, S., Mougharbel, I.: Comparative study of the optimization theory for cancer treatment. In: International Conference on Biomedical Engineering and Informatics (BMEI), vol. 2, pp. 927–933. IEEE (2011)
14. MeiHua, W., Su, X., Liu, F., Cai, R.: A cancer classification method based on association rules. In: International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 1094–1098. IEEE (2012)
15. Yao, L., Chung, Y.Y.: Mining cancer data with discrete particle swarm optimization and rule pruning. In: International Symposium on IT in Medicine and Education (ITME), vol. 2, pp. 31–34. IEEE (2011)
16. Kumar, D.A., Agarwal, V., Khandagre, Y.: Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support. In: CSI Sixth International Conference on Software Engineering (CONSEG), pp. 1–6. IEEE (2012)
17. Zibakhsh, A., Saniee Abadeh, M.: Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. *Engineering Applications of Artificial Intelligence* 26(4), 1274–1281 (2013)
18. Jesmin, N., Imam, T., Tickle, K.S., Chen, Y.-P.P.: Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications* 40(4), 1086–1093 (2013)
19. Ying, L., Wang, G., Chen, H., Shi, L., Qin, L.: An ant colony optimization based dimension reduction method for high-dimensional datasets. *Journal of Bionic Engineering* 10(2), 231–241 (2013)
20. Defeng, W., Shi, L., Heng, P.A.: Automatic detection of breast cancers in mammograms using structured support vector machines. *Neurocomputing* 72(13), 3296–3302 (2009)
21. Hualong, Y., Ni, J., Zhao, J.: ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing* 101, 309–318 (2013)
22. <http://www.breastcancerdeadline2020.org/about-the-deadline/progress-reports/2012-progress-report.html>
23. Jacques, F., Shin, H.-R., Bray, F., Forman, D., Mathers, C., Parkin, D.M.: Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer* 127(12), 2893–2917 (2010)
24. Cancer Australia. Report to the nation - breast cancer 2012, Cancer Australia, Surry Hills, NSW (2012)
25. Howlader, N., Noone, A.M., Krapcho, M., Neyman, N., Aminou, R., Altekruse, S.F., Kosary, C.L., et al.: SEER cancer statistics review(vintage 2009 populations), pp. 1975–2009. National Cancer Institute, Bethesda (2012)
26. Lawrence, G., Kearins, O., Lagord, C., Cheung, S., Sidhu, J., Sagar, C.: The Second All Breast Cancer Report. London: West Midlands Cancer Intelligence Unit. NCIN (2011)
27. Silverstein, M.J.S.M., Wyatt, J., Weber, G., Moore, R., Halpern, E., Kopans, D.B., Hughes, K.: Predicting the survival of patients with breast carcinoma using tumor size. *Cancer* 95(4), 713–723 (2002)
28. Australian Institute of Health and Welfare, Australian Association of Cancer Registries & National Breast and Ovarian Cancer Centre. Breast cancer survival by size and nodal status in Australia. Cancer series no. 39. Cat. no. CAN 34. Canberra: AIHW (2007)

29. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
30. Wang, Z., Sun, X., Zhang, D.: A PSO-based classification rule mining algorithm. In: Huang, D.-S., Heutte, L., Loog, M. (eds.) ICIC 2007. LNCS (LNAI), vol. 4682, pp. 377–384. Springer, Heidelberg (2007)
31. Behrouz, M.-B., Barmaki, R., Nasiri, M.: Mining numerical association rules via multi-objective genetic algorithms. *Information Sciences* 233, 15–24 (2013)