Breast Cancer Detection by Using Data Mining, a Review Study

Samia M. Gharib[1], Nevine Makram Labib[2], Rania A. Hodhod[3]
[1, 2] Computer and Information Systems Department, Sadat Academy for Management Sciences, Cairo, Egypt
[3]Columbus State University, Columbus, Georgia
sam4gharib@gmail.com, nevmakram@gmail.com, hodhod_rania@columbusstate.edu

Abstract:

Cancer is considered the second lethal disease in the world, with estimated 9.6 million deaths in 2018. Early detection of cancer can increase the survival rate and decrease both treatment costs and patients suffering. At the national level, this can reduce total annual economic cost of healthcare expenditure and loss of productivity. Predictive analytics using data mining and machine learning techniques have proven successful for early detection of cancer, identification of patients with high risk of survival, cancer morbidity, and mortality rate and predicting drug response. The aim of this survey paper is to review the important role of data mining and machine learning techniques in the detection of cancer. The paper provides a comparison between the most popular predictive tools and techniques, types of data, extracted features, error rate, diagnosis, associated factors, and estimation methods.

Keywords:

Data Mining, Predictive analytics, Machine learning, Knowledge discovery, Knowledge extraction, Cancer prediction

1. Introduction

According to 2018 statistics, breast cancer is the most dominant type of cancer in females worldwide [1]. Breast cancer comprises almost a quarter of new cancer cases among women; nearly one twentieth of women are expected to be diagnosed with breast cancer during their lifetime. Mortality rates in less developed countries are high though incidence rates are the lowest because of lack of access to early detection and treatment. Early detection can help to detect patients in early stages as well as those who have not been diagnosed yet [1]. In 2020, Female breast cancer was considered the major cause of global cancer incidence. It was the most commonly diagnosed cancer, with an estimated 2.3 million new cases (11.7%) of all cancer cases. It was estimated as the fifth leading cause of cancer mortality worldwide, with 685,000 total number of deaths [2].

In Egypt, breast cancer is the most prevalent cancer among women; it constitutes 29% of The National Cancer Institute's cases [3]. It comes in the second rank after liver cancer in new incidents and mortality rates. Number of new cases of breast cancer in Egypt in 2018, both sexes and; all ages was 23,081 people with a rate of 17.9% of the total number of new cancer cases as shown in Figure 1a. The total number of new female cases in 2018 for all ages was 23,081 women with a rate of 35.1% as shown in Figure 1b [4].

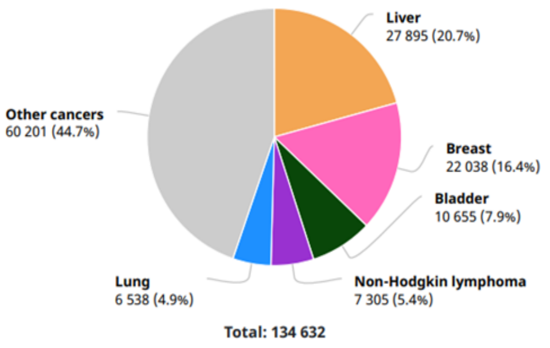

Figure 1a. Number of New Cancer Cases, Both Genders, Egypt, 2020 [4]

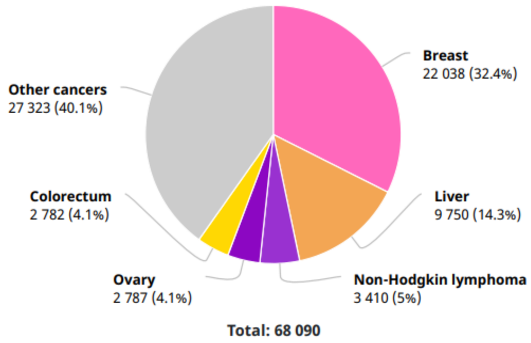

Figure 1b. Number of New Cancer Cases in Females, Egypt, 2020 [4]

1.1 Breast Cancer Stages

The stage of breast cancer is one of the most important factors in evaluating treatment options by the doctors. Breast cancer has 4 stages that can be summarized as follows [5]:

Stage 0: The tumors aren›t expanded to the neighboring tissues of the breast.

Stage 1: If the tumor is small (2cm size), but not expanded to lymph glands.

Stage 2 is divided into 2a, 2b. In Stage 2a, 2cm size, small tumor is spread to the lymph glands and tumor from 20mm (2cm) to 50mm (5cm) is not expanded to the axillary lymph glands. In Stage 2b: Tumor is greater than 5cm or tumor from 2cm to greater than 5cm then spread into the one or three axillary lymph glands.

Stage 3 is split into 3a, 3b, and 3c. Stage 3a: Tumor greater than 5cm expanded to 5 to 10 axillary lymph glands that are knitted with each other or with the neighboring tissues. Stage 3b: tumor expanded to the breast wall, skin or internal lymph glands. In this stage, it is not expanded to the remaining components of the body. Stage 3c: Tumor with any size that has expanded to more lymph glands.

Stage 4: any size of the tumor has been expanded to the remaining organs such as the distant lymph nodes, bones, brain, lungs, liver and chest wall.

1.2 Breast Cancer Early Detection & Diagnosis

Early detection of breast cancer is important because many women with breast cancer may have no symptoms. Early detection of breast cancer and getting the latest treatment is one of the important strategies to avoid deaths from breast cancer. Breast cancer in its early stages is small and not spread, which increases the success of the treatment. Performing regular screening tests is the most reliable method for early detection of breast cancer [6]. Breast Cancer can be early detected and diagnosed by the physicians using the means in Table 1.

Table1. Methods of Detecting & Diagnosing Breast Cancer [6]

Detect Breast cancer	Imaging Tests	Mammograms
		Breast Ultrasound
		Breast Magnetic Resonance Imaging (MRI)
	Biopsy	Breast Biopsy
Detect Breast Cancer Spread	Imaging Test	Chest X-Ray
		Computed tomography (CT) Scan
		MRI
		Ultrasound
		Positron emission Tomography (PET) Scan
		Bone Scan
Detect Breast Cancer During Pregnancy	Using mammograms and other imaging tests is up to the physician as exposing the developing fetus to radiation may be harmful, especially during the first trimester.	

2. Data Mining Techniques Used for Cancer Prediction

2.1 Data Mining Overview

Data mining is a logical process to search through large amounts of data for the purpose of finding useful data. The main objective of data mining is to find hidden patterns that were previously unknown. Uncovering these patterns help to make useful decisions [7]. Data mining has many applications in fraud detection, market analysis, scientific control, and medical domain…, etc. [8]. Data mining and machine learning techniques are commonly used for detecting breast cancer in early stages, which can help to decrease mortality rate among breast cancer diseases [17]. These are explained briefly in the next subsections.

2.2 Data Mining Techniques

2.2.1 Classification

Classification is the most commonly applied data mining technique; it uses a set of pre-classified examples to develop a model that can be used to classify the population of records at large. This approach is usually used in fraud detection and risk analysis. It uses decision tree or neural network-based classification algorithms. Data classification process involves 2 phases: learning and classification. In the learning phase, the training data are analyzed by a classification algorithm. In the classification phase, test data is used to evaluate the accuracy of the classification model. If the accuracy is acceptable, the model can be applied to new data input. Types of classification techniques that can be used include decision tree, Bayesian Classification, neural networks, support vector machines (SVM) and association rules [7].

2.2.2 Clustering

Clustering is the grouping of a particular set of objects based on their characteristics, assembling them by their similarities. This approach partitions the data by using a specific algorithm [9]. Types of clustering methods include partitioning methods, hierarchical agglomerative (divisive) methods, density based methods, grid-based methods, and model-based methods

2.2.3 Prediction

Regression techniques can be applied for prediction. Regression analysis is used to model the relationship between one or more independent variables and dependent variables. Independent variables are attributes already known and response variables are what we want to predict. Real World problems are more complex, so more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary for predicting future values. Types of regression methods include linear regression, nonlinear regression, multivariate linear regression, and multivariate nonlinear regression.

2.2.4 Association rules

Association rules are used to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases, such as relational databases, transactional databases, and other forms of data repositories. The types of association rules include multilevel association rule, multidimensional association rule, and weighted association rule.

2.2.5 Neural Networks

Neural networks can be defined as «a model of reasoning based on the human brain» [10]. It can be considered as the most common data mining technique, because it is a simple model of brain neural interconnections, adapted for use on digital computers. It learns from a training set, generalizing patterns inside it for classification and prediction. It can also be applied in undirected data mining and time-series prediction [11].

3. Related Work

Researchers in [12] proposed a new breast cancer predictive model, where they apply Weighted Naïve Bayes data mining technique to improve the accuracy of the prediction by using a benchmark dataset. The results show that the proposed prediction model was easy to use, readable, efficient, modifiable, interoperable, and can be used as a decision support tool for the physician.

In [13], during the examination of tested tissues to diagnose breast cancer, pathologists use

10 features to determine whether the tumor is malignant or not. Weka, ZeroR and decision trees were to predict the output feature; "class" which determined whether the tumor is benign or malignant.

In [14], the researchers applied classification by using C4.5 and Naïve Bayes algorithm. The results showed accuracy of 98.0966% from C4.5 and 95.8523% from Naïve Bayes algorithm, which show that the performance of C4.5 algorithm is better than Naïve Bayes algorithm. Regarding preprocessing, they eliminate non-cancerous data, such as attributes related to social factor, color, racial, and geographical condition. The objective of [15] was to predict breast cancer. The researchers applied three different common data mining techniques; Naïve Bayes, Radial basis function (RBF) Network, J48, in developing the prediction models. The results (based on indicated that Naïve Bayes was the best with an accuracy rate 97.36%. Naive Bayes was applied because of its simplicity, and classification algorithm is because it is a simple yet powerful model and it returns not only the prediction but also the degree of certainty

In [16], the researchers applied two prediction models of breast cancer by using 2 common data mining techniques; Naive Bayes and J48. They compared both data mining techniques by using performance factors classification accuracy and time of execution. The results showed that the Naïve Bayes was more accurate, with less execution time than J48.

The researchers in [17] proposed an efficient Breast Cancer Diagnosis (BCD) model to detect breast cancer by using a support vector machine (SVM) with 10-fold cross validation. The complexity of the problem increases if there are many input features used to diagnose breast cancer. Principal Component Analysis (PCA) was used to reduce the feature space from a higher dimension to a lower dimension. Results showed that the PCA increased the accuracy of the model. The proposed BCD model is compared with other supervised learning algorithms like decision trees (DT), random forest, k- Nearest Neighbors (K-NN), stochastic gradient Descent (SGD), AdaBoost, neural networks (NN), and Naïve Bayes. Evaluation parameters like F1 measure, receiver operating characteristic (ROC) curve, accuracy, lift curve and calibration plot showed that the proposed BCD model outperforms the other algorithms.

The framework proposed by [18] provided an extensible breast cancer prognosis framework (XBPF). Consists of susceptibility, recurrence, and survivability of breast cancer. A representative feature subset selection (RFSS) algorithm was suggested to be used with SVM in order to improve efficiency of prognosis. SEER dataset was used along with a prototype which was built to demonstrate proof of the concept. The results disclosed that the framework is useful in prognosis of breast cancer instead of focusing on susceptibility, survivability and recurrence. Individually, SVM-RFSS has shown significant performance improvement over other prognosis methods.

An interactive data analysis and visualization tool is proposed [19]. Visualizations were intended to compare the performance of three machine learning algorithms applied on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. This model includes two stages: input stage and analysis/visualization stage. The input stage allowed the user to upload his/her own data by means of the application user interface. During the second stage, the uploaded user data is analyzed and automatically visualized. If the user changed one of the parameters› values, displayed results will automatically be updated subsequently. The main objective of the model is to enable the user to interactively compare the performance of three different Machine Learning (ML) algorithms: KNN, SVM and Normal Bayes (NB) by using metrics of accuracy, sensitivity and error rate in an easy to use method. Supervised vector machine (SVM) classifier proved its efficiency and is considered the best classifier with the highest accuracy rate, compared to the other two. The classifiers› parameters could be adjusted by the user to get a maximum accuracy rate. Results revealed that all the selected algorithms showed high performance in predicting whether the cancer lump is benign or malignant. SVM showed the best results with 97.85% accuracy rate.

The researcher's objective in [20] is to develop a fully automatic method by adapting descriptor features extracted by a deep convolutional neural network (DCNN) for training and predicting of breast cancer in addition to pooling operation for the classification of hematoxylin and eosin stain (H&E) histological breast cancer images. Different data augmentation methods were used for the optimization of DCNN performance. The proposed network model showed 92.50% average classification accuracy.

Table 2, outlines the used data, data mining techniques, features , evaluation techniques, objective and accuracy rate of each previous literature study. From this table it can be seen that many researchers compared two or more data mining techniques as in [13], [14], [15] and [16]. Naïve Bayes data mining technique was used alone or with other techniques as in [12], [14], [15], and [16]. Researchers used Support Vector Machine (SVM) with other techniques in [17], [18], and [19]. Regarding the results, the confusion matrix was used alone or with other evaluation techniques as in [13], [15], [17], and [19]. Other evaluation techniques were used, like benchmark data set in [12], visual evaluation in [16], and comparative analysis in [20]. Concerning data sets used, Wisconsin data set was used in [17], and [19]. The Surveillance Epidemiology and End Results (SEER) data set was used in [14] and [18]. UC Irvine data set was used in [12] and [15]. From the accuracy rate we find that integrating Support Vector Machine (SVM) with Representative Feature Subset Selection (RFSS) achieved the highest accuracy rate 98.9% by using the Surveillance Epidemiology and End Results (SEER) dataset, which most of the studies used.

One of the most common limitations obvious in the studies included in this review is applying Machine Learning prediction techniques using very small sized datasets. This doesn›t provide sufficient data for training the ML model, which could lead to misleading predictions. Moreover, sampled data in the provided datasets belong to other geographic areas than Egypt, which means different incidence, Mortality rates and risk factors. Most of the surveyed studies applied evaluation and external/ internal validation, and only one researcher had performed Experiments on benchmark dataset. External validation enables more accurate and reliable predictions which minimize training bias.

4. Conclusion

Breast cancer is prevailing in women and is considered to be the second cause of death among them. Early detection of breast cancer helps women by increasing survivability. Most of the studies focused on classification Techniques, in addition to other techniques to increase the accuracy of the results. Most of these studies discuss in detail the application of breast cancer prediction techniques, but didn't shed the light on the preprocessing stage. Quality of the used dataset is very important for the accuracy of the results. The accuracy rates of these studies range from 60% to more than 90%. Extending these tools to real life applications can contribute greatly to the decrease of mortality rates caused by diagnosis of breast cancer in its late stages. Applying data mining in predicting breast cancer is essential to Egypt digital transformation, and sustainable development. The results from the existing literature show that data mining and predictive analytics can help in adopting inclusive healthcare coverage, improving healthcare service quality provision, and enhancing preventive and health programs.

References

[1] Torre, et al. (2016). Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 25(1), 16-27. https://doi.org/10.11589965-1055/.EPI-150578-

[2] Sung, H., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209.249-

[3] Omar, et al. 2003, Breast cancer in Egypt: a review of disease presentation and detection strategies. EMHJ - Eastern Mediterranean Health Journal, 9

[4] Global Cancer Observatory (GLOBOCAN) (2020). «Estimated number of cancer cases in 2020, Egypt" (http://gco. iarc.fr/). Accessed on: 1 March 2022

[5] Shailaja, K., et al. (2018). Prediction of breast cancer using big data analytics. Int J Eng Technol, 7(46), 223

[6] American Cancer Society (2017) Breast Cancer Early Detection and Diagnosis, [https://www.cancer.org/content/dam/CRC/PDF/ Public/8579.00.pdf]. Accessed on: 1 January 2020

[7] Bharati, M. & Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering.

[8] Ioni□□, I., & Ioni□□, L. (2016). Applying data mining techniques in healthcare. Stud Inform Control, 25(3), 385.94-

[9] M. Porkizhi (2017), A Study of Data Mining Techniques and Its Applications, International Journal for Science and Advance Research in Technology (IJSART).

[10] M. Negnevitsky (2002), Artificial Intelligence, a Guide to Intelligent Systems, England: Pearson Education Limited.

[11] Nevine M. Labib and and Michael N. Malek (2007, Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering,

[12] Shweta K and S.Soni (2016), Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection, International Journal of Computer Applications, 133.

[13] G. Sumalatha and S. Archana (2017), A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering.

[14] Yeulkar, K.(2017) ,Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R.

[15] Chaurasia V, Pal S. and Tiwari BB (2018), Prediction of benign and malignant breast cancer using data mining techniques, Journal of Algorithms & Computational Technology.

[16] C. Meera and D. Nalini (2018), Breast cancer prediction system using Data mining methods, International Journal of Pure and Applied Mathematics.

[17] Priyanka Israni (2019), Breast Cancer Diagnosis (BCD) Model Using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[18] R. Aavula and R. Bhramaramba (2019), XBPF: An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability, International Journal of Engineering and Advanced Technology (IJEAT).

[19] RA. Sanyour and M. Abdullah (2019), Real time data analysis and visualization for the breast cancer disease, Periodicals of Engineering and Natural Sciences.

[20] Kassani, et al. (2019). Breast cancer diagnosis with transfer learning and global pooling. In 2019 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 519524-). IEEE.