IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 18, Issue 4 Ser. 20 (April. 2019), PP 85-94
www.iosrjournals.org

Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review

Saria Eltalhi[1], Huda Kutrani [2]
[1](Computer Science Department, Faculty of Education, University of Benghazi, Libya)
[2](Health Informatics Department, Faculty of Public Health, University of Benghazi, Libya)
Corresponding Author: Saria Eltalhi

Abstract: *Breast cancer is the second cause of death among women. Early prediction of breast cancer will help with the survival of breast cancer patients. Data mining and machine learning have been widely used in the diagnosis of breast cancer and on the early detection of breast cancer. The aim of this research is to review the role of machine learning and data mining techniques in breast cancer detection and diagnosis. Most of these studies concentrated on diagnoses and prognoses breast cancer using WEKA tool. Most of the studies compared different classification algorithms to breast cancer prediction such as Decision tree, Naïve Bayes, and Artificial Neural Network.*

Keywords: *breast cancer, machine learning, data mining, classification algorithms, clustering algorithms*

Date of Submission: 18-04-2019 Date of acceptance: 04-05-2019

I. Introduction

In both developed and developing countries, Breast cancer is the most common cancer in women. Also, it is the second main cause of cancer death in women [1,2]. According to the WHO report in 2013, *"It is estimated that worldwide over 508 000 women died in 2011 due to breast cancer"*[3:3]. However, with early diagnosis, 97% of women could survive for five years or more [1,2].

Data mining and machine learning have been widely used in the diagnosis and prognosis of breast cancer [4]. Moreover, data mining and machine learning assist the medical researchers to identify relationships among variables and make them able to predict the outcome of disease using the historical datasets [4, 5]. Machine learning can be applied to improve breast cancer detection and diagnosis, as well as prevent over-treatment. Also, it could be an assistance to accurate decision making [4-7]. Therefore, the aim of this research is to review the role of machine learning and data mining techniques in breast cancer detection and diagnosis.

This research is organized as follows, Section II introduces a brief of breast cancer. Section III explains the algorithms and tools of data mining and machine learning used for breast cancer prediction. Section IV summarizes recent studies related to breast cancer diagnosis and prediction. Section V discusses the literature survey. Finally, Section VI concludes the research with a future scope.

II. Breast Cancer

2.1 Breast Cancer

Breast cancer occurs in breast cells, the fatty tissue or the fibrous connective tissue within the breast. Breast cancer is malignant tumors tend to become progressively worse and grow fast leading to death [1]. Although breast cancer is more common in females, it can rarely occur in males [1,2]. A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous) [2]. Factors such as age and a family history of breast cancer can increase the risk of breast cancer [2,8].

2.2 Types and stages of breast cancer

The treatment options for breast cancer are based on cancer stage and type [9,10]. The status of the cell surface receptors determines breast cancer classification [10].

Two types of tumors are identified [2,10,11]:

• ***Benign*****:** this tumor type is not dangerous for a human body and rarely causes human death. In this type, the tumor grows in one part (spot) of the body and has limited growth.

• ***Malignant*****:** this tumor type is more dangerous and causes human death, it is called breast cancer. The malignant tumor develops when cells in the breast tissue abnormally grow. The main types of breast cancer include:

1- Ductal carcinoma in situ (DCIS): is the earliest form of breast cancer and is curable.

2- Invasive Ductal Carcinoma (IDC): begin in the milk duct and is the most common breast cancer.
3- Invasive Lobular Carcinoma (ILC): start in a lobule of the breast. It has the ability to spread fast to the lymph nodes and other areas of the body.

The main stages of breast cancer [9-11]
Breast cancer stage is a description of tumor size and determines whether cancer has spread. The main stages of breast cancer include:

Stage I: Primary cancer, the tumor is small (two cm or less), and has not spread to lymph nodes.
Stage II: This stage is divided into two stages: stage IIA and stage IIB.
Stage IIA: The cancer is not larger than two centimeters but has spread to the lymph nodes under the arm (the axillary lymph nodes). OR the cancer is between two and five centimeters but has not spread to the lymph nodes under the arm.
Stage IIB: The cancer is between two and five centimeters, and has spread to the lymph nodes under the arm. OR the cancer is larger than five centimeters but has not spread to the lymph nodes under the arm.

Stage III & IV is known as 'Advanced Breast cancer'
Stage III: The tumors are bigger in size. They may have spread to lymph nodes, and to the surrounding breast tissue.
Stage IV: Tumors have spread to other parts of the body such as bone and lungs.

2.3 Treatment of breast cancer
Sometimes, patients treated with one of the treatments or combination of the treatments which are based on the woman's age, type and stage of cancer. The main treatments for breast cancer are:[2,8,10]
1- Surgery: There are two main types of surgery for breast cancer. The first type of surgery is called breast-conserving surgery or a lumpectomy. The goal of surgery is to remove the part of the breast containing cancer and some surrounding normal tissue. The second type of surgery is a mastectomy in which the entire breast is removed.
2- Radiotherapy: It kills cancer cells using gamma radiation.
3-Chemotherapy: It may use cytotoxic drugs to kill cancer cells both in the breast and elsewhere in the body.
4-Hormone therapy: It is often used after surgery to assist in reducing the risk of cancer coming back, or treat cancer that has spread to other parts of the body. It is usually taken for at least five years.
5- Biological therapy: It consists of new drugs that work differently from chemotherapy. It reduces the risk of breast cancer coming back.

2.4 Prevalence and survival rates of breast cancer
One million females are diagnosed with breast cancer approximately every year worldwide; therefore, it is the most common cancer in females [10,12]. Moreover, females' death due to breast cancer is high, 211,000 females in developed countries and 213,000 females in developing countries died from it in 2011 [12]. Sadly, new diagnosis and death from breast cancer are more incidents occurring in young females [12,13].

Survival prediction from breast cancer often uses a '5-year survival rate' after diagnosis, presented by percentage of females who are alive five years after the start of their treatment or diagnosis [10,12,13]. The survival rate of breast cancer is high in the early stage of the disease; 81% of females with early-stage breast cancer possibly survive for five years. However, only 35% of females with late or advanced stage breast cancer survive for five years [10,12,13].

III. Data Mining and Machine Learning

3.1 Data mining and machine learning

Data mining is the process of discovering interesting, meaningful patterns to represent knowledge from big data sets [6,7]. This knowledge provides useful information to improve decision support, prevention, diagnosis and treatment in diseases [2,14]. Data mining that can handle large volumes of data with multiple attributes, is defined as a logical process of discovering interesting patterns from huge data [15]. The purposes of data mining are knowledge discovery of patterns, reducing complexity, and saving processing time [15,16].

Machine learning is a learning program from past experience to improve its performance without human instruction (1) (2). There are two types of machine learning as following (2): Supervised Learning and Unsupervised Learning. Choosing one of Machine learning algorithm generally depends on data types and structures [17].

3.2 Data mining Algorithms

There are many algorithms such as Decision Trees, Naïve Bayes, k-means, and Neural Network; they are used for analyzing a huge amount of data. Some popular data mining algorithms are discussed in the following:

Decision tree algorithms (J48)

Decision tree algorithms are successful machine learning classification techniques. They are the supervised learning methods which use information gained and pruned to improve results. Moreover, decision tree algorithms are commonly used for classification in many research; for example, in the medicine area and health issues. There are many kinds of decision tree algorithms such as ID3 and C4.5. However, J48 is the most popular decision tree algorithm. J48 is the implementation of an improved version of C4.5 and is an extension of ID3 [6,7].

Regression

Regression analysis is a statistical methodology that is most often used for numeric prediction. There are various forms of regression, such as linear regression, multiple linear regression, logistic regression, and weighted regression. Also, there are several non-linear regression methods that are used for more complicated data analysis. Indeed, any regression technique whether linear or nonlinear can be used for classification. Regression also encompasses the identification of distribution trends based on the available data. Moreover, regression often yields good results in practice [6,7].

Association rules

Association rules are basically if/then statements which help us to find out the relationships between apparently unrelated data. Therefore, they can predict any of the attributes, not just a specified class. Indeed, association rules are generated from frequent patterns, and we can search for strong associations between frequent patterns to be used for classification. Moreover, association rules are very useful for analyzing and predicting patterns behavior [6,7].

Multilayer Perceptron

A multilayer perceptron is a simple two-layer neural network with no hidden layers. In fact, a two-layer perceptron (the input layer excluded) is adequate. The multilayer perceptron has the same significant power as a decision tree. The multilayer perceptron is an accurate predictor for the underlying classification problem. Indeed, learning a multilayer perceptron is closely related to logistic regression. However, multilayer perceptron has the advantage that they can learn to ignore irrelevant attributes. Recent studies have shown that multilayer perceptron can compete with more modern learning techniques on many practical datasets [6,7].

K-nearest-neighbours (kNN) algorithm

It is a simple supervised learning algorithm in pattern recognition. It is one of the most popular neighborhood classifiers due to its simplicity and efficiency in the field of machine learning [18].

KNN algorithm stores all cases and classifies new cases based on similarity measures; it searches the pattern space for the k training tuples that are closest to the unknown tuples. The performance depends on the optimal number of neighbors (k) chosen, which is different from one data sample to another [19].

Support Vector Machine (SVM)

It is a supervised learning method derived from statistical learning theory for the classification of both linear and nonlinear data. SVM classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separating [6,20].

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) defined as a model of reasoning based on how the human brain works [4,5]. ANN became the subject of active research over the past few decades, and it has been employed by more and more researchers. It consists of layers that are interconnected input, hidden and output layers. Neural network receives the data at the input layer and then processing the data by a hidden layer and providing the training results to the output layer [4]. The Network learns by adjusting the weights in the learning phase to be able to predict the correct class label of the input [5].

Naïve Bayes (NB)

It is a probabilistic classifier; it is one of the most efficient classification algorithms based on applying Bayes' theorem with strong (naïve) independent assumptions [9,21]. It assumes the value of the feature is independent of the value of any other features, given the class variable. Based on the maximum probability. It detects the class membership for the given tuple to a particular class [21].

3.3 Data mining tools

Data mining tools provide ready to use an implementation of the mining algorithms. Most of them are free open-source softwares. Some of the popular data mining tools are discussed in the following:

WEKA

The Weka is a collection of machine learning algorithms and data preprocessing tools for Knowledge Learning. The name (WEKA) stands for Waikato Environment for Knowledge Analysis. It is a computer program that was developed at the University of Waikato in New Zealand. The program is written in Java, and it runs on almost any operating system. It is a free data mining software. WEKA supports evaluating, visualizing, and preparing the input data. It also supports different learning algorithms such as classification, clustering, and regression [7].

Tanagra

Tanagra is a free machine learning software for research and academic purposes. It was developed by Ricco Rakotomalala at the Lumière University Lyon 2, France. Tanagra supports several data mining tasks such as visualization, descriptive statistics, regression, clustering, classification and association rule learning [22].

Orange

Orange is a Python-based tool for machine learning and data mining. Its visual programming interface is clean and easily understood. The orange may be more suited for novice researchers and small projects [23,24].

Matlab

Matlab as a data mining tool has an interpreted language and graphical user interfaces (GUIs). It also has hundreds of mathematical functions to support multi-paradigm numerical calculations which make it suitable to the computing environment [25,26].

C++

Programming language for general-purpose, it also has many useful features such as templates and reusability, it has object-oriented features with modular objects which could be tested independently [27].

IV. Literature survey

Twenty-four recent research articles have been reviewed to explore the computational methods to predict breast cancer. The summaries of them are presented below.

Chaurasia et al. [28] developed prediction models of benign and malignant breast cancer. Wisconsin breast cancer data set was used. The dataset contained 699 instances, two classes (malignant and benign), and nine integer-valued clinical attributes such as uniformity of cell size. The researchers removed the 16 instances with missing values from the data set to become the data set of 683 instances. The benign were 458 (65.5%) and malignant were 241 (34.5%). The experiment was analyzed by the Waikato Environment for Knowledge Analysis (WEKA). Naive Bayes, RBF Network, and J48 are the three most popular data mining algorithms were used to develop the prediction models. The researchers used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The models' performance evaluation was presented based on the methods' effectiveness and accuracy. Experimental results showed that the Naive Bayes had gained the best performance with a classification accuracy of 97.36%; followed by RBF Network with a classification accuracy of 96.77% and the J48 was the third with a classification accuracy of 93.41%. In addition, the researchers conducted sensitivity analysis and specificity analysis of the three algorithms to gain insight into the relative contribution of the independent variables to predict survival. The sensitivity results indicated that the prognosis factor ''Class'' was by far the most important predictor.

Yue et al. [4] reviewed studies on machine learning (ML) techniques applied in breast cancer diagnosis and prognosis. The researchers focused on studies using artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbors (k-NNs) techniques. And they also used the Wisconsin breast cancer database. Machine learning techniques have shown their remarkable ability to improve classification and prediction accuracy. The researchers provided a clear and intuitive catalog of information. This information was shown in a table with references, algorithms, sampling strategies and classification accuracy. The researchers believed that many algorithms have achieved very high accuracy using the Wisconsin breast cancer database (WBCD), but the development of improved algorithms is still necessary. In the future, the researchers intend to conduct an in-depth study of Friedreich Ataxia (FRDA) dataset using machine learning techniques to the purpose of establishing an intelligent FRDA healthcare system.

Banu & Ponniah [21] presented a classifier technique for breast cancer prediction. The researcher confirmed how Bayes classifiers like Tree Augmented Naive Bayes (TAN), Boosted Augmented Naive Bayes (BAN) and Bayes Belief Network (BBN) could be used for producing the best performance regarding classification and accuracy. The dataset used here is Wisconsin Diagnostic Breast Cancer (WDBC) which contains about 569 instances with 32 attributes. In order to improve the accuracy, all classifiers are combined with Gradient Boosting (GB) technique. Before applying GB process, the three classifiers had produced almost similar accuracy of 90.1%. However, the results of all the classifiers are enhanced with GB. The performance evaluation of the classifiers was done based on the prediction of accuracy, specificity, and sensitivity. The obtained results provide clear evidence of the benefits of TAN usage in breast cancer classification.

Akinsola et al. [29] presented a breast cancer prediction system that can aid doctors in predicting breast cancer status based on the clinical data of patients (classes were a benign and malignant tumor). The Home dataset from Federal government hospital in Lagos was used; it contained over 1700 instances. Eleven attributes such as cell size, cell shape, and the predicted class were selected. Three supervised learning algorithms were used to classify breast cancer data. C4.5, Multilayer Perceptron (MLP) and Naïve Bayes were investigated using WEKA toolkit. The performance evaluation of the three algorithms was done based on the prediction Accuracy and Time Taken to build the model. The C4.5 was the best model with the highest accuracy of 93.9%, and time taken was 0.28 second. The researchers suggested that the system needs to add another function to able patients from using the system.

Mirajkar and Lakshmi [30] predicted the type of cancer using Naïve Bayes Classification algorithm of data mining. Proposed method aimed to predict the risk of certain types of cancer. Based on the Naïve Bayes algorithm, symptoms of the cancer were classified to recognize the risk of cancer such as breast and ovarian.

Oskouei et al. [14] reviewed a comprehensive survey about all studies that applied data mining techniques in breast cancer diagnosis, treatment & prognosis. And presenting the main problems of these studies that still exist in this area. Forty-five articles were investigated; these articles divided into four categories based on the main goal. Twenty-one articles were compared to the accuracy of applying various classification techniques to diagnose breast cancers. Twelve articles proposed an approach to distinguish benign from malignant breast tumors. One of the articles applied Regression Data Mining Techniques for a diagnosis of the first stages of breast cancers in breast cancer databases. Eleven articles developed a prediction model for breast cancer (early diagnosis model or predicting breast cancer survival model). Based on the results of these studies, most of the studies were concerned with comparing the accuracy rate of data mining techniques. Unfortunately, there is no tool that automatically diagnoses breast cancer or proposes the best treatment for patients. The researchers recommended that attempt to develop a tool to apply data mining tools with the capability to automatically diagnose breast cancer and could propose the best treatment.

Mandal [31] the proposed system aims to find the smallest subset of features that can guarantee highly accurate classification. Before classification data preprocessing was done includes data cleaning, data dimensionality reduction, data transformation. The study used WDBC breast cancer dataset, and it is extracted from the UCI machine learning repository. The dataset contains 569 instances and 32 attributes which 70% of the instances were used for training purpose and 30% of the instances for testing purpose. The testing data were applied to three classification methods viz. Naïve Bayes (NB), Logistic Regression (LR), and Decision Tree (DT) classifiers were applied to the testing data which detect whether the cell is benign or malignant. The results showed that the Logistic Regression Classifier provided the best classifier with the highest accuracy with a reduced subset of features (four) and the time complexity of this algorithm is less than the other two classifiers.

Sumalatha & Archana [32] studied different data mining techniques for early diagnosis and prediction of breast cancer. The research work analyses the J48 and ZeroR algorithms to predict breast cancer. These two algorithms were applied using WEKA. Total instances of ZeroR analysis were 699. The three major steps used in this research, the collection of datasets, data preprocessing and classification.

Devi et al. [33] investigated automated diagnosis of breast cancer based on a machine learning algorithm. The proposed approach was a three steps process. In the first step, the data were grouped into a number of clusters using the Farthest First clustering algorithm. Due to shrinking the size of the dataset, the computation time reduced greatly. In the second step, outliers are detected in breast cancer dataset using ODA (Outlier Detection Algorithm). The third step identifies whether the cancer is benign or malignant in the pre-processed data set using J48 classification algorithm. Wisconsin Breast Cancer Dataset (WBCD) and Wisconsin Diagnosis Breast Cancer (WDBC) was used to test the efficacy of the proposed system. The experiments were performed using WEKA (Waikato Environment for Knowledge Analysis) version 3.7.13. Experimental results proved that the two steps proposed approach serves to be the best compared to the existing research for the same data set. The highest accuracy was 99.9% for WBCD data set and 99.6% for WDBC data set. This research will help the doctors to diagnose breast cancer and thereby helping the patients in recovery.

Padmapriya and Velmurugan [34] evaluated the performance of classification algorithms to analyze breast cancer data by analyzing the mammogram images based on its characteristics. Different attribute values of breast cancer affected mammogram images were considered for analysis in this work. Patients food habits, the age of the patients, their lifestyles, and occupation attributes were considered. Three popular data mining methods J48, AD-Tree, and CART was used for the analysis by using the WEKA tool. The performance evaluation of the three algorithms models was conducted based on prediction accuracy, specificity, sensitivity and kappa statistics. The study observed that the classifiers J48 had an accuracy of 98.1 %, AD-Tree with 97.7%, and the highest accuracy of 98.5 % is performed by CART. Moreover, the CART algorithm performed well for classifying mammogram images. The researchers recommended that other classification algorithms should be used for the analysis of the same mammogram images to predict their performances.

Abed et al. [35] developed a hybrid classification algorithm which uses the Genetic Algorithm (GA) and k Nearest neighbor algorithm (kNN). To help breast cancer physicians in early diagnosis for prediction. The data mining techniques were used to help breast cancer physicians to early diagnosis of breast cancer. The Primary purpose of the GA algorithm is to optimize techniques; it was used for kNN by selecting best features as well as the optimization of the k value, while the kNN is used for classification purpose. The dataset was obtained from the Wisconsin Breast Cancer Dataset extracted from the UCI Repository of Machine Learning. The researchers had performed experiments on WBCD were contained 699 instances and 10 attributes; whereas, WDBC consisted of 569 instances and 31 attributes. When compared the result of the proposed algorithm with different classifier algorithms; the researchers obtained better results from the proposed algorithm. The accuracy of the proposed approach was 99%.

Chidambaranathan [36] used a hybrid algorithm of k-means and ELM to predict breast cancer. The k-means algorithm is responsible for clustering tumors based on the extracted features. Each cluster represents a specific tumor pattern. ELM was extended to the generalized SLFNs which effectively classifies with greater detection accuracy in a lesser amount of time. A hybrid algorithm of k-means and ELM is retained the extracted features as input after that the image is classified with SVM as normal, benign or malignant. The specificity, sensitivity, jaccord distance, and accuracy are calculated. Results show that the proposed system works better than the others to predict breast cancer.

Lavanya et al. [37] presented breast cancer prediction system based on a hybrid approach; classification and regression trees (CART) classifier with feature selection and bagging technique for higher classification accuracy and improved diagnosis. They used the hybrid approach to enhance the classification accuracy of breast cancer and Feature Selection to remove irrelevant attributes that do not play any role in the classification task. The Bagging means Bootstrap aggregation was used to classify the data with good accuracy. Data were collected from machine learning repository of UCI where experiment three breast cancer datasets (Breast Cancer, Breast Cancer Wisconsin (original), Breast Cancer Wisconsin (diagnostic). The Breast Cancer Dataset contained 286 Instances and 10 Attributes; the Original Dataset contained 699 Instances and 11 Attributes. While the Diagnostic Dataset contained 569 Instances and 32 Attributes, all previews dataset with two classes.

Padmapriya and Velmurugan [38] predicted breast cancer by analyzing the mammogram images based on its characteristics. Researchers used the dataset of 250 patients with either malignant or benign type of tumor; the dataset was obtained from the Cancer Institute in India. Nine significant attributes were used. The data were recorded in the Excel data sheet file; the file was saved in the format of CSV which was converted into ARFF format to be accepted in WEKA software. The breast cancer data were classified based on patients' age and type of cancer (malignant or benign tumor). Researchers used J48, CART, and ADTree classification algorithms. The performance evaluation of the classification algorithms was based on the TP Rate, FT Rate, and precision analysis. The researchers found that the CART algorithm performs well at classifying mammogram images with 98.5 % of accuracy, followed by J48 classifiers with an accuracy of 98.1 %. The researchers recommended repeating the experiment to predict the performances of the other classification algorithms in analyses the same mammogram images.

Sivakami [39] proposed breast cancer Hybrid Model which integrates DT and SVM algorithms. This model was used to classify patients into two classes (Benign/Malignant). The dataset containing eleven attributes was obtained from Wisconsin Breast Cancer Dataset (WBCD) taken from UCI machine learning repository which contains 699 instances where 241 cases belong to the malignant class and 458 cases belong to the benign class. Sixteen instances of the dataset have missing values. The result was compared to IBL, SMO, and NAÏVE classifications techniques using Weka software. The results show that DT+SVM perform well in classifying the breast cancer data, better than any other classifier algorithms. The accuracy of the Classification model was DT – SVM 91%. The low error rate was 2.58%, correctly classified instance was 459 and incorrectly classified instance were 240.

Zand et al. [5] presented a comparative survey on data mining techniques in the diagnosis and prediction of breast cancer. And analysis of the prediction to find the most suitable technique for predicting cancer survival rate. Three classification techniques were used; Naïve Bayes, neural network, and C4.5

algorithms were investigated using WEKA toolkit. SEER breast cancer data set was used. The researchers pre-processed the input data set from the SEER database. They found almost half of the patients' records were missing data from the Extent of Disease attribute and Site-Specific Surgery attribute. Thus, these attributes were removed from the data set. Sixteen important clinical attributes such as age, tumor size, and node size were selected. The data-set contained 151,886 instances. The performance evaluation of the three algorithms was done based on prediction Accuracy, Precision, and Recall. The researchers observed that the accuracy of the prognosis analysis of the three classification techniques was highly acceptable. Also, it can help medical professionals in decision making for early diagnosis and avoid a biopsy.

Majal et al. [40] presented a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. The FP algorithm was used association Rule Mining approach to find the frequent patterns for the diagnosis of breast cancer type (benign and malignant). The researchers also used the Decision Tree algorithm in the classification approach was used to predict the prognosis of breast cancer based on three predictor attributes. The three attributes were age, gender, and intensity of symptoms to achieve a goal attribute (disease) which can be predicted from symptoms. Wisconsin data set was used; it contained 699 records and nine attributes. The researchers found that the accuracy of the diagnosis analysis is highly acceptable and can help the medical professionals in decision making to predict early diagnosis and avoid a biopsy.

Chaurasia and Pal [16] investigated the performance of different classification data mining techniques to develop accurate prediction models for breast cancer. The Wisconsin dataset from UCI machine learning was used. Ten attributes such as uniformity of cell size and the predicted class were selected. The diagnosis was classified as a malignant or benign tumor. The dataset contained 699 instances. Sixteen of instances with missing values were removed from the dataset to construct a new dataset with 683 instances. Three supervised learning algorithms were used to classify breast cancer data. IBK, BF Tree and Sequential Minimal Optimization (SMO) were investigated using the WEKA machine learning environment. The performance evaluation of the three algorithms was done based on the prediction accuracy of the model. The comparison results showed Sequential Minimal Optimization (SMO) has higher prediction accuracy (96.2%) than IBK and BF Tree methods. Moreover, SMO had 0.92 of Kappa statistic (KS) and less of mean absolute error (MAE) than IBK and BF Tree methods. All attributes were important to breast cancer survival. These attributes were tested using Chi-square test, Info Gain test, and Gain Ratio test.

Joshi J. et al. [41] this research used four cluster algorithms to early diagnoses of breast cancer patients. The dataset was obtained from the UCI web data repository. Recordset with 10 attributes (age, Menopause, Tumor-size, inv-nodes, Node-caps, etc.). Clustering techniques such as K-Means, Hierarchical Cluster Method (HCM), Expectation Maximization and Farthest First were applied. The research used a WEKA tool and applied different data mining algorithm to measure accuracy and performance to detect breast cancer. This research work used Farthest First (FF) algorithm to diagnose a patient's health. The clustered instances were 286 instances; 219 (77 %) of them were healthy and 67 (23%) of them were sick. This FF technique resulted in two clusters comprising of healthy, and sick diagnosis of breast cancer patients. The experimental work using EM algorithm generated 117(41%) healthy, 65 (23%) sick and 104 (36%) No Class. Applying the HCM algorithm, the outcome was 285 (99.65 %) healthy and 1 (0.35%) sick. When research work used k-means algorithm, the experimental work result was 236 (83%) healthy and 50 (17%) sick. k-means clustering algorithm and FF algorithm yielded an accuracy of 83% and 77% respectively which were helpful to early diagnosis of the breast cancer patients.

The results as shown below in Table 1

Clustering Technique	Healthy	Sick	NO Class
FF	77	23	-
EM	41	23	36
HCM	99.65	0.35	-
k-means	83	17	-

Sumbaly et al. [42] proposed a predictive model for early detection of breast cancer using data mining techniques, namely the J48 decision tree algorithm. The paper discussed various data mining technique for breast cancer diagnosis and also summarized breast cancer (types, risk factors, symptoms, and treatment). The dataset used was obtained from the Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository. The dataset contained 699 Instances of breast cancer patients and 11 Attributes. Over 200 kinds of known cancers were classified by the type of cell that is affected. The research was focused on Breast cancer. Breast cancer occurs in both male and female, but it is more common among women. Since the causes of breast cancer are not identified there are many women with breast cancer who have no apparent risk factors. Decision tree creates a predictive model. The tree consists of leaves and branches, where leaves represent the class labels and branches represent conjunctions of a feature leading to the class labels. Pre-processing is required when

applying J48 where two attributes of the dataset were changed. The outcome was 661 (94.5637 %) correctly Classified and 38 (5.4363%) wrongly Classified. When compared the result of the proposed algorithm with different classifier algorithms; the results showed effectiveness the proposed model (J48 classifiers) with feature selection was a superior technique that could be applied to breast cancer diagnosis.

Joshi et al. [43] tried to develop a novel prototype of a clinical problem concerning diagnosing and managing patients with breast cancer. Ten important clinical attributes such as age, tumor size, and node size were selected. Web Usage Mining data was used to find the hidden patterns in the breast cancer dataset. Data were collected from UCI dataset repository. The experiment was analyzed by WEKA Open Source environment; 37 classification rules were used to diagnose and predict breast cancer among patients. The class of diagnosis was healthy or sick patients. Experimental results showed that more the accurate result was obtained from 13 of the 37 classifiers with a diagnosis; 76% were healthy and 24% were sick. The thirteen classifiers were Bayes-Net, SMO, Logistic, Multilayer-Perceptron, J48, SGD, Simple-Logistic, AdaBoostM1, Attribute Selected, Filtered-Classifier, and Classification via Regression, Multi-Class Classifier, and LMT. The researchers recommended that repeated the experiment to make a prototype for predicting the best classifier using Tanagara and Orange data mining tool.

Chandrasekar et al. [44] studied breast cancer prediction using data mining techniques. The study aimed to develop accurate prediction models for breast cancer with a neural network classification technique. An ensemble approach was used for possible improvements. The classification techniques included Lazy IBK, Tree Random Forest, Lazy K Star classifier, and Rules NNge were applied. Data were collected from the WBCD dataset. The experiment was analyzed by WEKA software. The dataset contained 286 instances which 201 of them were benign and 85 were malignant. These instances were described by 10 Attributes such as age, tumor size, and class. In conclusion, Tree Random classifier achieved a classification accuracy of 98%. The researchers proposed using to analyze Ensemble classifier for 100% accuracy.

Gupta et al. [45] presented an overview of studies using data mining techniques to predict the diagnosis of breast cancer. Eighteen articles were investigated; these articles were divided into two categories based on the main goal. Ten articles studied data mining classification techniques for breast cancer diagnosis. Eight articles studied data mining classification techniques for breast cancer prognosis. Based on the results of these studies, data mining techniques offer great promise to discover patterns hidden in the data that can help doctors in decision making. Also, the accuracy of the diagnosis analysis of various applied data mining classification techniques is highly acceptable. The prognostic problem is mainly analyzed under ANNs and its accuracy came higher in comparison to the other classification techniques applied for the same. The best model was obtained after building several different types of models, or by trying different technologies and algorithms.

Bellaachia and Guven [46] presented an analysis of the prediction of the survival rate of breast cancer patients using data mining techniques. Three classification techniques were used; Naïve Bayes, back-propagated neural network and C4.5 decision tree algorithms were investigated using WEKA software. SEER breast cancer data set was used. The researchers developed a set of tools to extract and clean-up the raw SEER data set. They found almost half of the patients' records were missing data from the Extent of Disease attribute and Site-Specific Surgery attribute. Therefore these attributes were removed from the data set. Sixteen important clinical attributes such as age, tumor size, and node size were selected. The dataset contained 151,886 instances. The researchers have divided the data set into two pre-classification processes; the first process was contained 23.2% of the data set were *"not survived"* and 76.8% of the data set were *"survived"*. The second process contained 58.3% of the data set were *"not survived"* and 41.7% of the data set were *"survived"*. The researchers considered the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD) for construction of survivability prediction models. The performance evaluation of the three algorithms models was done based on prediction Accuracy, Precision, and Recall for each process. The study observed that the C4.5 algorithm had a much better performance than the other two techniques.

V. Discussion

This research summarizes some of the recent studies were done in data mining concerning breast cancer. Data mining algorithms can be effectively used to ‘mine’ relevant information extracted from the huge amounts of data generated by healthcare services. These studies showed that the results are better when applying many of algorithms rather than applying a single algorithm on a data set. WEKA is chosen in most of the research. C++, Tanagra, Matlab, etc. are some of the other popular tools used for data analysis. Careful selection of the algorithms combination and accurate implementation of them on the data set give an effective implementation of diagnoses and prognoses breast cancer's system. The required dataset is divided into two parts; one is used for learning of algorithms and the smaller partition is used for verifying. Most of the studies were a comparison of different classification techniques on a dataset to correctly classify if a given patient has benign or malignant breast cancer. Other studies have applied a model of predicting breast cancer survival, or

the breast cancer risk factor model. Commonly used classification techniques are Decision tree, Naïve Bayes, Artificial Neural Network, Association Rule, Support Vector Machine, and Regression.

VI. Conclusion

Breast cancer is the most common cancer in women and the second main cause of cancer death in women. When the early symptoms of breast cancer are ignored, the patient might end up with drastic consequences in her health and can lead to death. Breast cancer can be kept under control when it is detected early. Many studies focus mainly on the application of classification techniques to breast cancer prediction; rather than studying various home data cleaning and pruning techniques that can prepare and make a dataset suitable for mining. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

References

[1]. World Health Organization. Cardiovascular diseases (CVDs). https://www.who.int/cardiovascular_diseases/en/. [Accessed 3rd January 2019].
[2]. Mayo Clinic. Breast Cancer: Symptoms and causes - [Internet]. Mayo Clinic. 2016. Available from: https://www.mayoclinic.org/diseases.../ breast-cancer/symptoms-causes/syc-20352470 [Accessed 5th January 2019].
[3]. World Health Orgaization. Breast cancer:prevention and control. WHO; report 2016.
[4]. Yue W, Wang Z, Chen H, Payne A, Liu X. "Machine learning with applications in breast cancer diagnosis and prognosis". Designs. 2018; 2(2):13.
[5]. Zand HK. "A comparative survey on data mining techniques for breast cancer diagnosis and prediction". Ind. J. Fundam. Appl. Life Sci. 2015; 5 (S1):4330-9.
[6]. Han J, Kamber M, Pei J. "Data mining: concepts and techniques". (3rd Ed.)2012; San Francisco, CA, USA: Morgan Kaufmann Publishers.
[7]. Witten IH, Frank E, Hall MA, Pal CJ. "Data Mining: Practical machine learning tools and techniques". (3rd Ed.)2011; San Francisco: Morgan Kaufmann.
[8]. NHS. Breast cancer in women: Treatment - NHS [Internet]. Available from: https://www.nhs.uk/conditions/breast-cancer/treatment/ [Accessed 8th January 2019].
[9]. Maughan KL, Lutterbie MA, Ham PS. Treatment of breast cancer. Am Fam Physician. 2010; 81(11):1339–46. DOI: 10.1002/1097-0142(19810501)47:9<218.
[10]. Roche. Breast cancer a guide for journalists on breast cancer and its treatment. p. 1–10.
[11]. Genentech. Types and Features of Breast Cancer. 2015; 1–2.
[12]. Anderson BO. G LOBAL B REAST C ANCER T RENDS UICC World Cancer Congress 2014 : Breast Heal Glob Initiat. 2014;1–10.
[13]. Shulman LN, Willett W, Sievers A, Knaul FM. "Breast cancer in developing countries: opportunities for improved survival". Journal of oncology. 2010; 2010: 595167. Available from doi:10.1155/2010/595167
[14]. Oskouei RJ, Kor NM, Maleki SA. "Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges". American journal of cancer research. 2017; 7(3):610-27.
[15]. Zand HK. "A comparative survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction". Indian Journal of Fundamental and Applied Life Sciences. 2015; 5(S1):4330-4339.
[16]. Chaurasia V, Pal S. "A Novel Approach for Breast Cancer Detection using Data Mining Techniques". (IJIRCCE) International Journal of Innovative Research in Computer and Communication Engineering.2014; 2(1): 2456-65
[17]. Lundin M, Lundin J, et al. "Artificial neural networks applied to survival prediction in breast cancer". Oncology. 1999; 57(4):281-6.
[18]. Jabbar MA, Deekshatulu BL, Chandra P. "Classification of heart disease using k-nearest neighbor and genetic algorithm". International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA).2013; 10:85-94.
[19]. Hassanat AB, Abbadi MA, Altarawneh GA, Alhasanat AA. "Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach". (IJCSIS) International Journal of Computer Science and Information Securit. 2014; 12(8):33-9
[20]. Srinivas R." Managing Large Data Sets Using Support Vector Machines". Computer Science and Engineering. M.Sc. Thesis, University of Nebraska - Lincoln, 2010
[21]. Banu B, Thirumalaikolundusubramanian P." Comparison of Bayes Classifiers for Breast Cancer Classification". Asian Pacific journal of cancer prevention (APJCP). 2018; 19(10):2917-20.DOI: 10.22034/APJCP.2018.19.10.2917
[22]. Naik A, Samant L. "Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime". International Conference on Computational Modeling and Security (CMS 2016). 2016; 85:662-8. doi: 10.1016/j.procs.2016.05.251
[23]. Jovic A, Brkic K, Bogunovic N. "An overview of free software tools for general data mining". 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2014: 1112-7.
[24]. Slater S, Joksimović S, Kovanovic V, Baker RS, Gasevic D. "Tools for educational data mining: A review". Journal of Educational and Behavioral Statistics. 2017; 42(1):85-106.
[25]. Amutha.T, Priya. M. "Data Mining in Cloud Usage Statistics with Matlab's Facts and Device Studying Toolbox". International Journal of Pure and Applied Mathematics.2018; 118(20): 901-908
[26]. Shaykahian GA, Martin DE, Beil R." Improve Data Mining and Knowledge Discovery through the use of MatLab". National Aeronautics and Space Administration (NASA). 2011
[27]. Kohavi R, Sommerfield D, Dougherty J. "Data mining using MLC++: A machine learning library in C++". International Journal on Artificial Intelligence Tools.1997; 6(4): 537-566
[28]. Chaurasia V, Pal S, Tiwari BB. "Prediction of benign and malignant breast cancer using data mining techniques". Journal of Algorithms & Computational Technology. 2018; 12(2):119-26. DOI: 10.1177/1748301818756225

[29]. Akinsola AF, Sokunbi MA, Onadokun IO. "Data Mining For Breast Cancer Classification". International Journal of Engineering And Computer Science. 2017; 6(8): 22250-22258. DOI: 10.18535/ijecs/v6i8.06
[30]. Mirajkar P, Lakshmi P. "Prediction of Cancer Risk in Perspective of Symptoms using Naïve Bayes Classifier". International Journal of Engineering Research in Computer Science and Engineering. 2017; 4(9):145-149.
[31]. Mandal SK. "Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree". International Journal of Engineering and Computer Science. 2017; 6(2):20388-91. DOI: 10.18535/ijecs/v6i2.40
[32]. Sumalatha G, Archana S. "A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques". International Journal of Innovative Research in Computer and Communication Engineering. 2017; 5(6):11045-11050. DOI: 10.15680/IJIRCCE.2017.
[33]. Devi RD, Devi MI. "Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer". International Journal of Advanced Engineering Technology.2016; VII (II): 93-98
[34]. Padmapriya B, Velmurugan T." Classification Algorithm Based Analysis of Breast Cancer Data". International Journal of Data Mining Techniques and Applications. 2016; 5(1):43-49.
[35]. Abed BM, et al. "A hybrid classification algorithm approach for breast cancer diagnosis". In2016 IEEE Industrial Electronics and Applications Conference (IEACon). 2016: 269-274.
[36]. Chidambaranathan S. "Breast Cancer Diagnosis Based on Feature Extraction by Hybrid of k-means and extreme learning machine algorithms". ARPN Journal of Engineering and Applied Sciences. 2016; 11(7):4581-86
[37]. Lavanya D, Rani KU. "Ensemble Decision Tree Classifier for Breast Cancer Data". International Journal of Information Technology Convergence and Services (IJITCS). 2016; 2(1):17-24. DOI: 10.5121/ijitcs.2012.2103
[38]. Padmapriya B, Velmurugan T. "Classification Algorithm Based Analysis of Breast Cancer Data". International Journal of Data Mining Techniques and Applications. 2016; 5(1):43-49.
[39]. Sivakami K. "Mining big data: Breast cancer prediction using DT-SVM Hybrid model". International Journal of Scientific Engineering and Applied Science (IJSEAS). 2015; 1(5):418-29.
[40]. Majali J, Niranjan R, Phatak V, Tadakhe O. "Data Mining Techniques for Diagnosis and Prognosis of Cancer". International Journal of Advanced Research in Computer and Communication Engineering. 2015; 4(3):613-16. DOI 10.17148/IJARCCE.2015.43147
[41]. Joshi J, Doshi R, Patel J. "Diagnosis of Breast Cancer Using Clustering Data mining Approach". International Journal of Computer Applications. 2014; 101(10):13-7.
[42]. Sumbaly R, Vishnusri N, Jeyalatha S. "Diagnosis of Breast Cancer Using Decision Tree Data Mining Technique". International Journal of Computer Applications. 2014; 98(10): 16-24.
[43]. Joshi J, Doshi R, Patel J. "Diagnosis and prognosis breast cancer using classification rules". International Journal of Engineering Research and General Science. 2014; 2(6):315-23.
[44]. Chandrasekar RM, Palaniammal V, Phil M. "Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis". IOSR Journal of Computer Engineering (IOSR-JCE). 2013; 15(5):39-44.
[45]. Gupta S, Kumar D, Sharma A. "Data mining classification techniques applied for breast cancer diagnosis and prognosis". Indian Journal of Computer Science and Engineering (IJCSE). 2011; 2(2):188-95.
[46]. Bellaachia A, Guven E. "Predicting breast cancer survivability using data mining techniques". J Am Aging Assoc. 2006; 58(13): 10-14.

Saria Eltalhi and Huda Kutrani. "Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review." IOSR Journal of Dental and Medical Sciences (IOSR-JDMS), vol. 18, no. 04, 2019, pp 85-94.