DOI : https://dx.doi.org/10.26808/rs.re.v3i5.04
International Journal of Research in Engineering & Science ISSN:(P) 2572-4274 (O) 2572-4304
Available online on http://rspublication.com/IJRES/IJRE.html volume 3 Issue 5 Sep-Oct 2019

A SURVEY ON THE DATA MINING TECHNIQUES USED FOR BREAST CANCER PREDICTIONS

Ishola, O. B.[1], Ajibola, A. A.[2]
[1]Department Of Computer Science, University Of Abuja, Nigeria
+234-9094261633, isholaolabisi@gmail.com
[2]School of Computing and Engineering, University of Huddersfield, UK
+44-7394-133652, a.showole@hud.ac.uk

ABSTRACT
Breast cancer has become a major cause of death for women in the world. The most effective way to reduce the rate of death caused by breast cancer is early detection. In the last few years, there has been an increase in the usage of data mining techniques on medical data, to discover useful patterns or trends that are used in analysis, diagnosis and decision making. Data mining algorithms, when used appropriately, are efficient in improving the quality of prediction, diagnosis and disease classification. In this paper we present an overview of the data mining techniques used for the classification of medical data and also highlight some related works in breast cancer predictions, using a table to compare the results obtained from the classifications.
Keywords: Breast cancer, data mining, medical data, predictions, classification

Corresponding author: Ishola O. B.

INTRODUCTION
According to WHO(2018), the second leading cause of death globally is cancer. It is responsible for an estimated 9.6 million deaths in 2018. Globally, about 1 in 6 deaths is due to cancer. Approximately 70% of deaths from cancer occur in low- and middle-income countries. Developing nations only enjoy 5% of global funds for cancer control and very few human and material resources are also available in such countries [1][2][3].
The American Cancer Society defines cancer as a group of diseases that is characterized by the unrestricted growth and spread of abnormal cells. The spread, if not controlled, can result in death[4]. Breast cancer is a type of cancer that affects the breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk[3][5]. All over the world, the rising breast cancer incidence and mortality represents a significant and growing threat for the developing world. Breast cancer is on the rise across developing nations, mainly due to the increase in life expectancy and lifestyle changes such as women having fewer children, as well as hormonal intervention such as post-menopausal hormonal therapy. [6]
Recently, data from different fields such as banking, retail, telecommunications, medical diagnostics, etc. include valuable information and knowledge which is often hidden. Processing these volumes of data and retrieving meaningful information from it is always a herculean task. Data Mining is a powerful tool for handling such tasks. The application of data mining techniques in breast cancer research has been one of the important research topics in medical science in recent years [2]. The classification of breast cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. There are many techniques to predict and classify breast cancer pattern. [6]

The objective of this study is to summarize the various techniques used in articles on breast cancer predictions. It gives an overview of the data mining techniques being used on breast cancer datasets for predictions.

A. Breast Cancer: An Overview

Breast cancer is a malignant tumor which develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division.[7][8]. The cells are the building blocks that make up all tissues and organs of the body, including the breast. Normal cells in the breast grow and divide to form new cells as they are needed. When normal cells grow old or get damaged, they die, and new cells take their place but sometimes, the process goes awry and the old or damaged cells don‘t die as they should. This forms a buildup of extra cells which results to the formation of massive tissues called a lump, growth, or tumor. Tumors in the breast can be benign (not cancerous) or malignant (cancerous)[6].

Although breast cancer is the most common cancer among women and the second leading cause of cancer death in women, but the survival rate is high, especially in developed countries. With early diagnosis, 97% of women survive for 5years or more [7][8].

Although scientists do not know the exact causes of most breast cancer, they know some of the risk factors that increase the likelihood of a woman developing breast cancer[8].

Risk Factors

Breast cancer is caused by a number of factors called risk factors; they are classified as either modifiable (those that can be controlled like habits, environmental hazards, etc) or non-modifiable factors (those that cannot be controlled like, gender, family history etc)[3]. Most women have more than one known risk factor for breast cancer, yet will never get the disease. The most common risk factors for breast cancer is not only being female and growing older. There may be more than one cause of breast cancer[9].

The risk factors include, but not limited to the following:

- being a woman
- getting older
- starting menstrual periods at an early age (before age 12)
- having an inherited mutation in the BRCA1 or BRCA2 breast cancer gene
- lobular carcinoma in situ (LCIS)
- a personal history of breast or ovarian cancer
- a family history of breast, ovarian or prostate cancer
- having high breast density on a mammogram
- having a previous biopsy showing atypical hyperplasia
- starting menopause after age 55
- never having children
- not breastfeeding
- having your first child after age 35
- alcohol intake
- smoking tobacco
- exposure to pesticides, chemicals and organic solvents
- radiation exposure, frequent X-rays in youth
- high bone density
- being overweight after menopause or gaining weight as an adult
- high natural levels of sex hormones
- use of oral contraceptives.
- postmenopausal hormone use (current or recent use) of estrogen or estrogen plus progestin[3][4][5][6][9][10][11][12].

DOI : https://dx.doi.org/10.26808/rs.re.v3i5.04
International Journal of Research in Engineering & Science ISSN:(P) 2572-4274 (O) 2572-4304
Available online on http://rspublication.com/IJRES/IJRE.html volume 3 Issue 5 Sep-Oct 2019

B. Knowledge Discovery In Databases (Kdd) And Data Mining

Recently, the volume of data being collected and stored in databases has increased due to the advancements of researchers' interest in the areas of machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization[9]. This section provides an introduction to knowledge discovery and data mining.

Knowledge Discovery in Databases(KDD)

Knowledge Discovery in Databases(KDD) and Data Mining are collaborative areas focusing upon methodologies for extracting useful knowledge and patterns from data. The two terms are often used interchangeably. The term, Knowledge Discovery in Databases or KDD for short, can be defined as the broad process of finding knowledge in data, and it emphasizes distinctive application of particular data mining methods. KDD also refers to the significant process of extracting implicit, hidden and potentially useful information from data in databases. Meanwhile, data mining is an important step in the KDD process[9][13]. Fig. 1 below shows data mining as a step in an iterative knowledge discovery process.

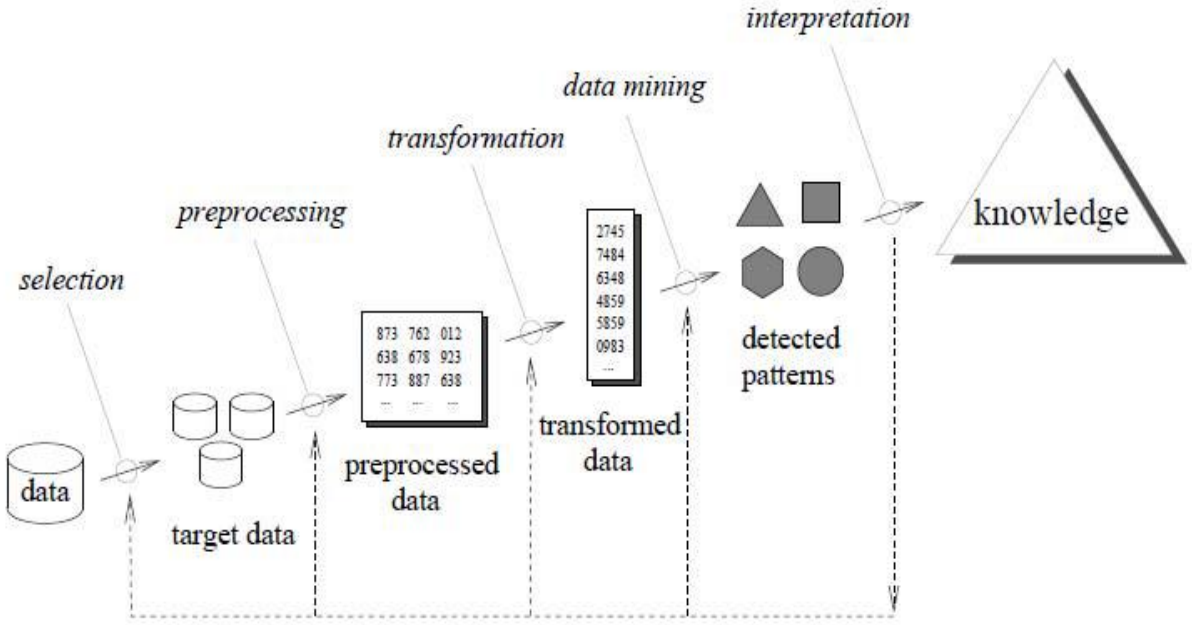

Figure 1: The Knowledge Discovery in Databases(KDD) process[13]

The Knowledge Discovery in Databases process is comprised of a few steps, starting from collections of raw data to their transformation to some form of new knowledge. The iterative process consists of the following steps:

(1) *Data cleaning(* also known as data cleansing): it is a phase where noise data, inconsistent and irrelevant data are removed from the collection.

(2) *Data integration*: this is the stage where multiple data sources, often heterogeneous, are combined in a common source.

(3) *Data selection:* this is the level in which the data that are relevant to the analysis are decided on and retrieved from the database.

(4) *Data transformation(* also known as data consolidation): this is the phase in which the selected data are transformed into forms appropriate for mining, by performing some summary or aggregation operations.

(5)Data mining: it is the crucial step in which clever techniques are applied to extract data patterns that are potentially hidden and useful.
(6) *Pattern evaluation:* here, strictly interesting patterns representing knowledge are determined based on some given measures.
(7) *Knowledge presentation:* is the final phase where the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results[13][14].

Data Mining Process

Data mining is the process of deriving patterns from data; the patterns that may be discovered depend on the data mining tasks that are performed on the dataset. There are two basic data mining tasks:

- descriptive data mining tasks, which help to understand the specific properties of the dataset
- predictive data mining tasks, which are used to perform predictions on the available dataset[3].

Data mining applications can use different criteria to examine data. These may include:

- association (patterns that define the interconnection of data and events),
- sequence or path analysis (patterns where one event leads to another),
- classification (identification of new patterns with predefined targets) and
- clustering (combining identical or similar objects together)[13].

The basic steps involved in mining data are:

- *Problem definition: This* is the definition of the goals and objectives and the identification of tools. It will make it easier to build the corresponding behavioural model.
- *Data exploration:* Quality of data must be verified for suitability to produce an accurate model. Hence recommendations on future data collection and storage strategies can be made.
- *Data preparation: This* is the process of cleaning and transforming data, in order to remove missing and invalid data. All known valid values are made consistent for a robust analysis.
- *Modeling: This* involves the use of data mining algorithms for the purpose of analysis, based on the data and the desired outcomes. The specific algorithm is selected based on the particular objective to be achieved and the quality of data to be analyzed.
- *Evaluation and deployment: This* is the analysis and interpretation of the results gotten from the analysis done by the data mining algorithms in order to create recommendations for consideration[3]. Fig. 2 shows the steps involved in the data mining process.

BACKGROUND

There are various papers that have been documented and published on the use of data mining techniques, statistical methods and machine learning algorithms that have been applied for breast cancer diagnosis and predictions. This section contains the review of some of those articles.

In [15], Delen et al. compared Artificial Neural Network(ANN), logistic regression and decision tree techniques for breast cancer survival analysis. They used the SEER data's

DOI : https://dx.doi.org/10.26808/rs.re.v3i5.04
International Journal of Research in Engineering & Science ISSN:(P) 2572-4274 (O) 2572-4304
Available online on http://rspublication.com/IJRES/IJRE.html volume 3 Issue 5 Sep-Oct 2019

twenty variables in their prediction models. The decision tree which had 93.6% accuracy and ANN 91.2%, were found more superior to logistic regression which had 89.2% accuracy.

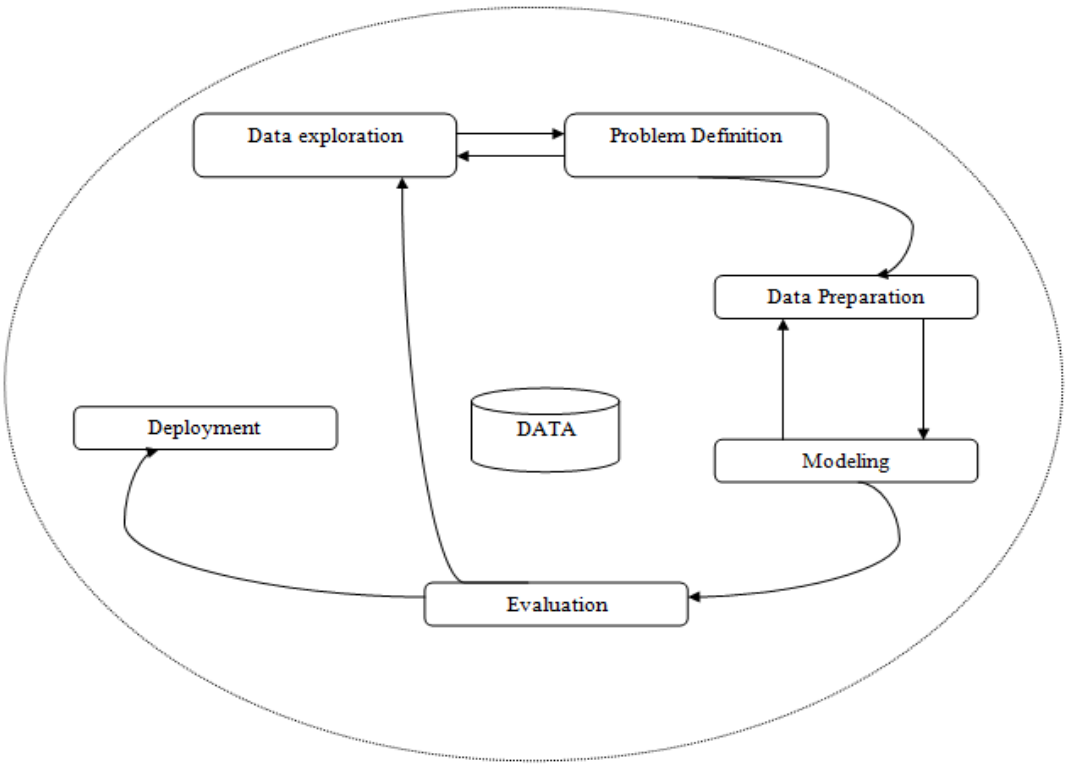

Figure 2: The Data Mining Process[3]

In [16], Choi et al compared the performance of an Artificial Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN(88.8%) and Hybrid Network(87.2%) were very similar and were better than the Bayesian Network. It was found out that the proposed Hybrid model can also be useful to take decisions[13].
In [17], Street applied Artificial Neural Network (ANN) classification to Wisconsin Prognostic Breast Cancer and SEER datasets for the analysis of survival. He established a novel encoding as good and poor prognosis of censored data in an ANN architecture to provide a framework for prognostic prediction.

In [18], Bellaachia and Gauven used C4.5 decision tree along with two other techniques, i.e. Naïve Bayes and Back-Propagated Neural Network. They used the new version of the SEER Breast Cancer Data to present an analysis of the prediction of survivability rate of breast cancer patients, using the three data mining techniques above. They have adopted a different approach in the pre-classification process by including three fields: STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death) and used the Weka toolkit to experiment with these algorithms. They found out that the model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques.

In [19], Abdelaal et al. examined the capability of the Support Vector Machine (SVM) classification with Tree Boost and Tree Forest in analyzing the Digital Database for Screening Mammography (DDSM) dataset for the extraction of the mammographic mass features, along

with age that discriminates true and false cases. Here, SVM technique was found to show a more promising result than the tree boost and tree forest, because it increased the diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve.

In [20], Chang and Liou investigated the artificial neural network, decision tree, logistic regression, and genetic algorithm. Comparative studies were done and the accuracy and positive predictive values for each algorithm were used as the evaluation indicators. WBC database was incorporated for the data analysis followed by the 10-fold cross-validation. The results showed that the genetic algorithm was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible. The genetic algorithm model yielded better results than other data mining models for the analysis, in terms of the overall accuracy of the patient classification, the expression and complexity of the classification rule.

In [21], Khan M.U. et al. studied a hybrid scheme based on fuzzy decision trees on SEER data and experiments were performed using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques. After comparing the performance of each for cancer prognosis, it was found that the hybrid fuzzy decision tree classification is more robust and balanced than the independently applied classification.

CLASSIFICATION ALGORITHMS

Data mining consists of various methods that serve different purposes, each method having its own advantages and disadvantages. However, most data mining methods commonly used for this review are of classification category and the applied prediction techniques usually assign patients to either a "*benign*" or *"don't have"* group (i.e. non- cancerous) or a "*malignant*" or *"has"* group (i.e. cancerous) and then rules are generated for the same. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems.

In data mining, classification is one of the most important task. It is a supervised learning in which targets are predefined and data are mapped in to the predefined targets. Classification is aimed at building a classifier, based on some cases with some attributes which describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification are described as follows[13]14].

A. Naïve Bayes' classifier

Naive Bayes Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data.[3]

Let X be a data sample containing instances, Xi where each instance is the breast cancer risk factors (modifiable and non-modifiable). Let H be a hypothesis that X belongs to class C which contains (unlikely, likely and benign cases). Classification is to determine P(Hj|X), (i.e., posteriori probability): the probability that the hypothesis, Hj (unlikely, benign or likely) holds given the observed data sample X.

Bayes' theorem is stated as

$$P(H_j|X) = (P(X|H_j) * P(H_j)) / P(X)$$

DOI : https://dx.doi.org/10.26808/rs.re.v3i5.04
International Journal of Research in Engineering & Science ISSN:(P) 2572-4274 (O) 2572-4304
Available online on http://rspublication.com/IJRES/IJRE.html volume 3 Issue 5 Sep-Oct 2019

where

- $P(H_j|X)$ is the probability of hypothesis H_j given the data X. This is called the posterior probability.
- $P(X|H_j$ is the probability of sample data X given that the hypothesis H_j was true.
- $P(H_j)$ is the probability of hypothesis H_j being true (regardless of the data). This is called the prior probability of H_j.
- P(X) is the probability of the sample data being observed (regardless of the hypothesis).[3]

B. Support Vector Machine(SVM)

Support vector machine (SVM) is a maximum margin classification algorithm that is established on the theory of statistical learning. It is a method for classifying both linear and non-linear data. It aims at finding a linear separator (hyper-plane) between the data points of two classes in multidimensional space, maximizing the margin separating both classes while minimizing the classification errors and makes use of a non-linear mapping technique to convert the original training data into a higher dimension one. SVMs are well suited to dealing with interactions among features and redundant features[13] [23].

C. Evolutionary Programming (EP) and Genetic Algorithms (GAs)

Evolutionary programming and Genetic algorithms are algorithmic optimization methods which are motivated by observing the principles in natural evolution. When there is a collection of potential solutions to problems which seem to compete with each other, the best solutions are selected and combined with each other. In doing so, it is expected that the overall goodness of the solution set will become better and better, as related to the process of evolution of a population of organisms. Evolutionary programming and genetic algorithms are used in data mining for the purpose of formulating hypotheses about dependencies between variables[13].

D. Neural Networks

Neural networks are enormous volume of interconnected nodes that perform summation and thresholding in loose comparison with the neurons of the brain. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units whereby each connection has a weight that is associated with it. The network learns in the learning phase by adjusting the weights in order to predict the correct class label of the input[23].

E. K Nearest Neighbours

K-Nearest Neighbor (KNN) is a classifier that classifies instances based on their similarity. Each instance is considered as a point in multi-dimensional space and classification is done based on the nearest neighbors. The value of 'k' for nearest neighbors is the number of cases that can be considered as neighbors, so as to determine how to classify an unknown instance. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. However, some disadvantages with nearest-neighbor classifier cannot be overlooked. The time taken to classify a test instance increases linearly with the number of training instances that are kept in the classifier, hence, it requires a large storage space. Also, there is a decrease in performance with increasing noise levels. It also performs badly when different

attributes effect the outcome to different extents. One parameter that can affect the performance of the algorithm is the number of nearest neighbors to be used, so it uses just one nearest neighbor by default.[6]

F. RBF Network

A radial basis function(RBF) network is an artificial neural network that uses radial basis functions as activation functions. RBFs were first introduced in the solution of the real multivariable interpolation problems. It was used in the design of neural networks. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks are used for varying purposes including classification and time series predictions[9]. RBF networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer, as shown in figure 3.

G. Simple Logistic

Logistic regression is a generalization of linear regression and a very powerful modeling tool used to assess the likelihood of a disease or health condition as a function of a risk factor (and covariates). Simple and multiple logistic regression both assess the association between independent variable(s) (Xi) -- sometimes referred to as exposure or predictor variables — and a dichotomous dependent variable (Y) – referred to as the outcome or response variable. It is used primarily for predicting binary or multiclass dependent variables[6].

H. Decision Trees (DT's)

A decision tree is a tree where each non-terminal node represents a test on an attribute, each link or branch represents an outcome or decision on the considered data item[13].
Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node- the topmost node- and follow the assertions down until we reach a terminal node (or leaf) which represents the outcome. A decision is made when a terminal node

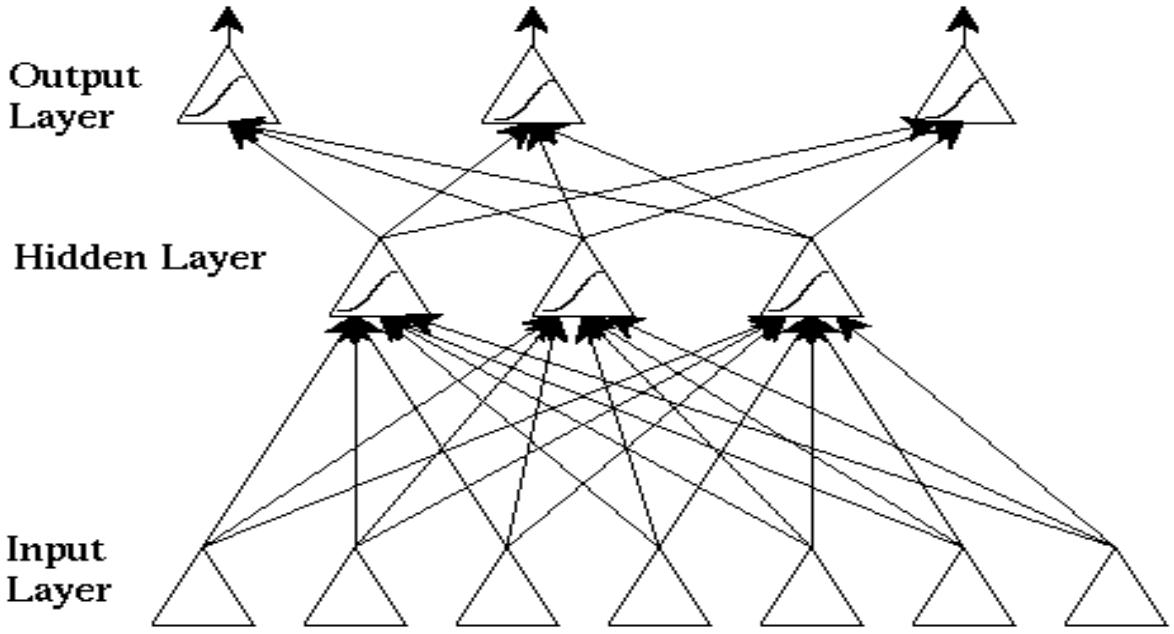

Figure 3: The structure of RBF neural network.[9]

DOI : https://dx.doi.org/10.26808/rs.re.v3i5.04
International Journal of Research in Engineering & Science ISSN:(P) 2572-4274 (O) 2572-4304
Available online on http://rspublication.com/IJRES/IJRE.html volume 3 Issue 5 Sep-Oct 2019

is approached. Decision trees can also be explained as a special system of a rule set, which are symbolized by their hierarchical organization of rules[3]. There are different types used for classification problems: ID3, C4.5 also known as J48, which is a successor to ID3, CART, CHAID, etc.[28].

MATERIALS AND METHODS

Datasets for breast cancer can be gotten directly from hospital databases or from online repositories. Some of the review papers got their datasets from hospitals while others were from online repositories (see Table 1). Each dataset has different attributes and varying number of attributes.

The UCI machine learning repository is located in breast-cancer-Wisconsin sub-directory, filenames root: breast-cancer-Wisconsin having 699 instances, and 9 integer-valued attributes. 16 instances with missing values were removed from the dataset during preprocessing, giving a new dataset with 683 instances.

The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) is a source of epidemiologic information on the incidence and survival rates of cancer in the United States. The SEER Public-Use Data is an online repository that can be used for breast cancer predictions. Due to missing values, the data are preprocessed to contain different records and attributes.

Table 1: Some works done on breast cancer prediction, with the techniques and attributes used.

Paper Ref. No	Algorithms used	Attributes	Dataset	Conclusion
3	J48, Naïve Bayes	Family History, Existence of Benign Breast disease, Mammographically Dense Breast, Age at First Birth, Age at Menopause, (BMI), Endogenous Estrogen Levels, Waist-Hip Ratio, Age, Smoking Frequency, Alchohol Intake, Occupational Hazard, Current Oral Contraceptive use, Breast Feeding.	Datasets of patients' information from LASUTH	J48 decision trees is a better model
6	Sequential Minimal Optimization (SMO), IBK(*K Nearest Neighbours classifier)* and Best First Tree methods.	Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses	Wisconsin dataset from UCI	SMO classifier had a best result with highest accuracy of 96.2% , low error rate and performance.
9	RepTree, Radial Basis Function (RBF) Network and Simple Logistic. All are decision tree techniques.	Age, Menopause, Tumor size, Inv-nodes, Node caps, Degree of malignancy, Breast(L or R), Breast quadrant, Irradiation	*Cancer database of University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia*	Simple logistic classifier had the best result with accuracy of 74.47%
15	ANN, logistic regression and decision tree techniques	Race, marital status, primary site code , histologic type, behavior code, grade, extension of tumor , lymph node involvement, site specific surgery code, radiation, stage of cancer, age, tumor size, number of positive nodes, number of nodes, number of primaries.	SEER data	The decision tree with 93.6% accuracy and ANN with 91.2% were found more superior to logistic regression with 89.2% accuracy.
16	Artifical Neural Network, a Bayesian Network and a Hybrid Network	Same as for Ref.15 above	SEER data	The accuracy of ANN(88.8%) and Hybrid Network(87.2%) were very similar and were

				better than the Bayesian Network.
18	Naïve Bayes, the back-propagated neural network, and C4.5 decision tree algorithms.	Same as for Ref.15 above	SEER data	C4.5 gave the highest accuracy of 86.7%
24	J48, CART, ADTree and BFTree	sex, age, present problem, past history, medical diagnosis, occupation, food habit, height and weight	Swamy Vivekananda Diagnostic Centre Hospital, Chennai- India	J48 classifier had the highest out of the four algorithms that were used.
25	Logistic Regression, Naive Bayes, LinearSVC, SVM with linear kernel and Random Forest	Same as for Ref.6 above	Wisconsin dataset from UCI	The highest accuracy of 97.8% was gotten from Random Forest Classifier.
26	J48, Random tree, Random Forest, REPTree, Priority based decision tree algorithm.	Same as for Ref.6 above	Wisconsin dataset from UCI	Priority based classifier classified with an accuracy of 93.63%
27	Artificial neural network (ANN), regression tree (C&RT), logistic regression, and Bayesian belief network (BBN).	Same as for Ref.15 above	SEER data	Bayesian network had the highest accuracy.

EVALUATION METHOD

Most of the papers in review used the WEKA toolkit to experiment with different combinations of data mining algorithms. All experiments described in the papers were performed using libraries from WEKA machine learning environment. The WEKA is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA can be used as a data mining tool to evaluate the performance and effectiveness of prediction models built from several techniques because it offers a well defined framework for experimenters and developers to build and evaluate their models.

CONCLUSION

In this paper, we outlined and discussed some research works done for diagnosis or predicting breast cancers, and some of the data mining algorithms and techniques being used, with different datasets. Also, a table is given, which compares the accuracy of classifiers with different attributes and the generated output. However, it is observed that the accuracies vary depending on the classifiers used and the data set.

In the future, it will be good to apply all the classifiers on one or different data sets in order to compare their performances, and then conclude which is the most optimal. An intelligent agent may also be developed which will automatically diagnose or predict breast cancer, given some input data.

REFERENCES

1. World Health Organization Report, 2018
2. Grey, N and Sener, S. (2006) Reducing the global cancer burden, http://www.hospitalmanagement.net/features/feature648/, Date accessed 21 November 2012.
3. K.Williams, P.A.Idowu, J. A. Balogun, and A. I. Oluwaranti, Breast cancer risk prediction using data mining classification techniques, *Transactions on Networks and Communications, 3*(2), 01, 2015.
4. American Cancer Society. Breast Cancer Facts & Figures 2018. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

5. J. Sariego, Breast cancer in the young patient, *The American surgeon* **76** (12): 1397–1401, 2010.
6. V. Chaurasia, and S. Pal, A novel approach for breast cancer detection using data mining techniques, *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1), 2014.
7. J.M.Jerez-Aragonés, J.A. Gómez-Ruiz, G.Ramos-Jiménez, J.Muñoz-Pérez, and E. Alba-Conejo, A combined neural network and decision trees model for prognosis of breast cancer relapse, *Artificial intelligence in medicine*, *27*(1), 45-63, 2003.
8. H.K.K. Zand, A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences*, *5*(s1), 4330-9, 2015.
9. V.Chaurasia, & S. Pal, Data mining techniques: To predict and resolve breast cancer survivability, *International Journal of Computer Science and Mobile Computing*, 3(1), pp. 10 – 22. 2014.
10. P. Boffetta, M. Hashibe, C. La Vecchia, W.Zatonski, and J.Rehm, The burden of cancer attributable to alcohol drinking. *International Journal of Cancer* 119 (4), pp.884–7, 2006.
11. K.C.Johnson, A.B.Miller, N.E.Collishaw, J.R.Palmer, S.K.Hammond, A.G. Salmon, K.P. Cantor, M.D.Miller, N.F.Boyd, J. Millar, and F. Turcotte, Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk, *Tobacco control* **20** (1), 2009.
12. R.Ferro, Pesticides and Breast Cancer, *Advances in Breast Cancer Research* 01 (03): pp.30–35,2012.
13. S.Gupta, D. Kumar, and A. Sharma, Data mining classification techniques applied for breast cancer diagnosis and prognosis, *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(2), 188-195, 2011.
14. J.Han and M.Kamber, Data Mining: Concepts and Techniques, 2nd ed., *San Francisco, Morgan Kauffmann Publishers*,2006.
15. D.Delen, G.Walker, and A.Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial intelligence in medicine*, *34*(2), pp.113-127,2005.
16. J.P.Choi, T.H.Han, and R.W. Park, A Hybrid Bayesian Network Model for Breast Cancer Prognosis, *Journal of Korean Society Med. Informatics*, 15(1), pp.49-57, 2009.
17. W.N.Street, A Neural Network Model for Prognostic Prediction, *Fifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann*, pp. 540-546, 1998.
18. A.Bellaachia, and E. Guven, Predicting breast cancer survivability using data mining techniques, *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining Age*, *58*(13), pp.10-110, 2006.
19. M.M.A.Abdelaal, H.A. Sena, M.W. Farouq, and A.B.M. Salem, Using data mining for assessing diagnosis of breast cancer, In *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference*,pp. 11-17, IEEE, 2010.
20. W.P.Chang, and D.M. Liou, Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data, *J Telemed Telecare*, *9*(1), 26, 2008.
21. M.U.Khan, J.P.Choi, H. Shin, and M. Kim, Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, In *Engineering in Medicine and Biology Society (EMBS) 30th Annual International Conference of the IEEE*, pp. 5148-5151, 2008.
22. A.E.Hassanien, and J.M. Ali, Rough set approach for generation of classification rules of breast cancer data, *Informatica*, *15*(1), pp.23-38, 2004.
23. L.G.Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M.Ebrahimi, and A.R. Razavi, Using three machine learning techniques for predicting breast cancer recurrence, *J Health Med Inform*, *4*(124), 3, 2013.
24. E.Venkatesan, and T. Velmurugan, Performance analysis of decision tree algorithms for breast cancer classification, *Indian Journal of Science and Technology*, *8*(29), 2015.
25. Ruolan Xu and Qiongjia Xu, Applying Different Machine Learning Models to Predict Breast Cancer Risk, 2017.

26. P.Hamsagayathri, and P.Sampath, Performance analysis of breast cancer classification using decision tree classifiers, *Int J Curr Pharm Res*, *9*(2), 19-25, 2017.
27. E.Y.Kibis, I.E.Büyüktahtakin, and Ali Dag, Data analytics approaches for breast cancer survivability: comparison of data mining methods, *Institute of Industrial and Systems Engineers (IISE) Annual Conference. Proceedings*, pp. 591-596, 2017.
28. https://www.linkedin.com/pulse/what-decision-trees-types-why-important-srinivas-kilambi.