JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Utilizing Cutting-Edge Machine Learning Methods for Breast Cancer Prediction

Pratibha Awasthi
(Research Scholar)
SATI Vidisha
Email: awasthipratibha170@gmail.com

Shaila Chugh
Department of Computer Science & Engineering
SATI Vidisha

Abstract: ***Background***: One of the illnesses that kill a lot of people each year worldwide is breast cancer. It can be difficult to identify and treat this kind of illness early on in order to lower the death toll. These days, a variety of machine learning and data mining approaches are employed in medical diagnostics, which has demonstrated its effectiveness in making predictions about chronic diseases like cancer that might potentially save the lives of those who suffer from them. Finding the prediction accuracy of classification algorithms such as Support Vector Machine, J48, Naïve Bayes, and Random Forest and suggesting the optimal approach is the main goal of this work.

Objective: This study aims to evaluate the efficiency and efficacy of the categorization algorithms in terms of prediction accuracy.

Methodology: Using the open-source WEKA tool, this paper applies a 10-fold cross-validation technique to the Wisconsin Diagnostic Breast Cancer dataset, analyzing the prediction accuracy of various classification algorithms, including Support Vector Machine, J48, Naïve Bayes, and Random Forest.

Findings(Results): According to the study's findings, Support Vector Machine has the best prediction accuracy, at 97-89%, and the lowest error rate, at 0.14%.

In Conclusion: This paper provides a clear view over the performance of the classification algo- rithms in terms of their predicting ability which provides a helping hand to the medical practition- ers to diagnose the chronic disease like breast cancer effectively.

Keywords: : Diagnosis, classification algorithms, Breast cancer, Machine Learning, Data mining, SVM, J48.

1. INTRODUCTION

The governments of all developing nations show their special commitment to the health care industry by setting aside funds for it in order to provide patients with easy-to-access, affordable care; but, in certain instances, the industry is lagging behind in the use of ICT [1,2]. Since deadly diseases like cancer are spreading quickly around the world and account for a large number of deaths each year, early detection of these illness indicators can save the lives of many people [3]. Cancer is one of the six most severe critical illnesses. Medical consultants can make more informed judgments based on the costs of treatment up front and reduce overall costs by utilizing ICT and a few IT technologies [2,4]. ICT stands for information and communication technology

These days, the most common illness identified in female bodies is breast cancer. Breast cancer is the second leading cause of death after lung cancer; yet, breast cancer is a disease that primarily affects women and is the cause of their deaths [5-8]. Approximately 16% of deaths worldwide are attributable to breast cancer.

The human body's cells react strangely, if not abnormally, during breast cancer sickness, and their properties change. Since the body is experiencing losses at this point quickly, it is crucial to identify these serious disorders as soon as possible [3, 7-8, 10-11]. When treating a serious illness like breast cancer, medical experts use one of two approaches [6]:

(i) This type of treatment involves localized applications of radiation therapy and surgery.

(ii) Chemotherapy and hormone therapy are examples of systematic treatments used in this type of care.

illnesses There are two stages of breast cancer presentation: benign (1) or malignant (2). A benign tumor is considered to be non-cancerous because it will not spread to other parts of the body, however at this stage the diagnosis cannot be made without surgery. A malignant tumor has the potential to spread to other parts of the body if it is not properly identified [9].

The old and traditional approaches are not beneficial due to the enormous amount of data [3,9,11]. In fact, we can claim that these methods do not give medical experts, who are the observers, with accuracy because many patterns, facts, and information are buried in the whole process. It requires microanalysis to achieve accuracy that is beyond the capabilities of a human observer. Therefore, technologies like machine learning, neural networks, and data mining are always helpful for the goal, and they give the physical sciences new hope. These cutting-edge methods offer prediction, prompt identification, reduction, and time and cost savings.

“Prediction of Breast Cancer”

Descriptive models and predictive models, often known as [14–17] and [12–14], are the two types of models used in data mining:

1.1. Characteristic Model

This model is based on the idea of unsupervised learning; to do this, data summarization, association rules, and clustering must be done.

1.2. Model of Prediction

This method, which is also known as supervised learning, is mainly concerned with predicting results using various types of analysis, such as regression and classification.

This work aims to predict the accuracy of different classifiers, such as Random Forest, Naïve Bayes, J48, and SVM, and to identify which classifier is the most efficient. These classifiers were chosen based on their popularity and ranking among the top 10 machine learning algorithms [18]. We used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to evaluate their performance.

The following sections make up this document.

(A) This section provides background information about breast cancer and a synopsis of the paper's contents.

(B) Part 2 explores the applicability of earlier research on supervised learning, data mining, and machine learning in predictive analysis.

(C) The study approach is described in Section 3, along with the dataset and classifiers that were selected.

(D) Test results are displayed in Section 4. comparing and contrasting the classifiers using different metrics and research instruments.

(E) The findings of the investigation are summarized in Section 5. Research.

2. CONNECTED WORKS

The inadequate nature of existing medical diagnoses, particularly for chronic disorders, was highlighted by Liou DM and Chang WP [20]. They recommended the use of a variety of machine learning approaches to enhance diagnosis and prognosis. Classifiers are crucial to predictive analytics, according to Leila Ghasem Ahmad, Abbas Toloie Eshlaghy, Alireza Poorebrahimi, Mandana Ebrahimi, and Amir R Razavi [21]. Using data from the Iranian Breast Cancer Center, they tested four methodologies: C4.5, decision trees, support vector machines (SVM), and artificial neural networks (ANN). SVM yielded a 95.7% accuracy rate. This discussion was aided by Saurabh Pal [22] and Vikas Chaurasia. To determine which classification methodology is best for detecting breast cancer, researchers compared a number of approaches. They assessed J48, RBF Neural Networks, SVM with RBF Kernel, Naïve Bayes, and Facilitating the timely detection and avoidance of long-term health issues.

Breast cancer comes second behind lung cancer in North America, making it one of the deadliest malignancies globally, according to Thoranin Intarajak and Seung Hwan Kang [5].

Vikas Chaurasia and Saurabh Pal [9] used the WEKA Tool to compare the accuracy rates of the Naïve Bayes, RBF, and J48 prediction models on the Wisconsin Breast Cancer (original) dataset. They were able to achieve 97.36%, 96.17%, and 93.41% percent accuracy, respectively.

Using the Wisconsin Breast Cancer (original) dataset, Thomas Noeld [12] examined the machine learning algorithms SVM, C4.5, Naïve Bayes, and K-NN. SVM achieved the maximum accuracy of 97.13%.

Godwin Ogbuabor and Ugwoke F.N.[13] assert that data mining methods and instruments are critical for locating hidden patterns and creating connections between various items.

3. METHODS USED

A dataset with several characteristics and labelled classes is often used in supervised learning to create a classification model. The Training Dataset and the Testing Dataset are the two primary components of the classification job [25]. The Testing Dataset is used to verify the model's performance, and the Training Dataset is used to build a predictive model incorporating the dataset's attributes. utilizing the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, we ran our investigation utilizing SVM (Support Vector Machine), J48, Naïve Bayes, and Random Forest as classifiers to predict cancer kind, namely Benign or Malignant. The 10-fold cross-validation method was applied in order to improve the precision of our findings.

3.1. Simple Bayes

This machine learning algorithm is frequently used for jobs involving categorization. It is predicated on Thomas Bayes' probability theory [9, 26]. It is well-known for being straightforward and efficient [18, 27–28], and it helps create classification models for predictions more quickly. The two components of this method are Bayes and Naïve. It is called "naïve" because it considers some traits to stand alone from others.

The term "Bayes" alludes to the statistician and philosopher Thomas Bayes, who is the subject of the theorem. In formal terms, the theorem says:

P (B|A)*P (A)/P (B) == P (A|B) (1)

2. J48

These models are frequently used for data inspection in the data mining industry. They depict a sort of classifier with a structure resembling a tree, in which a node [29] may act as a decision node designating a test to be run on a single-valued attribute or as a leaf node showing the value of the target attribute or class. J48 has adaptability while managing both continuous and categorical variables, which makes it easier to build an extensive decision tree. Furthermore, it has the capacity to handle missing values. With J48, pessimistic pruning may be used to remove superfluous branches from the tree, improving classification [9, 30].

3.3. SVM

Support vector machine, or SVM for short, is a supervised learning technique used extensively in data analysis, pattern recognition,

To ascertain the most well-liked class for a certain input x, a voting procedure is carried out among the trees on a random selection of the data [30–31]. An upper bound on generalization error in a random forest is derived using two factors [31, 33].

(a) Individual Classifier Accuracy

(b).dependence between different classifiers.

3.5. Concerning the Dataset

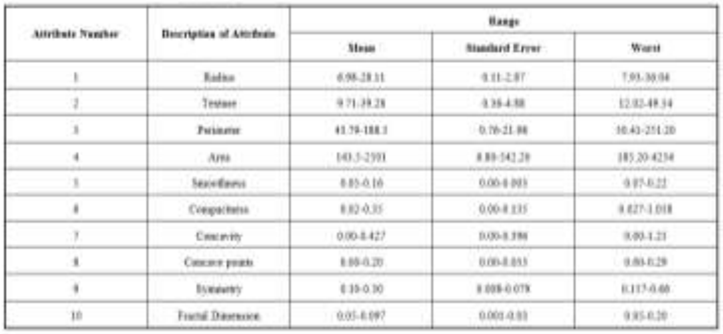
We used the Wisconsin Diagnostic Breast Cancer dataset (WDBC) [19] for this study. Many academics have used this dataset extensively [34–36]. There are 569 cases total, 357 of which are categorized as benign and 212 as malignant. A patient ID, thirty attributes related to the tumor diagnosis, and one class property designating the tumor diagnosis (malignant or benign) are all included in each instance. Information on the is provided in Table Number 1.

4. EXPERIENCE AND OUTCOMES

The classifiers used in our study are thoroughly analyzed in this part, with an emphasis on how well they worked to provide findings that were more accurate. For result validation, we applied the 10-fold cross-validation method. Version 3.8 of the popular open-source data mining application WEKA, which supports a number of *machine learning methods, was* utilized to get the findings.

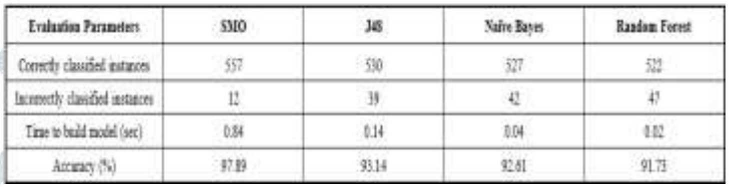
(a) The accuracy %, properly classified instances, wrongly classified instances, and the time needed to create the

Table 1. Details of the Attribute [18, 24].



Attribute Number	Description of Attribute	Range		
		Mean	Standard Error	Worst
1	Radius	6.98-28.11	0.11-2.87	7.93-36.04
2	Texture	9.71-39.28	0.36-4.88	12.02-49.54
3	Perimeter	43.79-188.5	0.76-21.98	50.41-251.20
4	Area	143.5-2501	6.80-542.20	185.20-4254
5	Smoothness	0.05-0.16	0.00-0.003	0.07-0.22
6	Compactness	0.02-0.35	0.00-0.135	0.027-1.058
7	Concavity	0.00-0.427	0.00-0.396	0.00-1.25
8	Concave points	0.00-0.20	0.00-0.053	0.00-0.29
9	Symmetry	0.10-0.30	0.008-0.079	0.157-0.66
10	Fractal Dimension	0.05-0.097	0.001-0.03	0.05-0.20

Table 2. Results for effectiveness.



Evaluation Parameters	SMO	J48	Naïve Bayes	Random Forest
Correctly classified instances	557	530	527	522
Incorrectly classified instances	12	39	42	47
Time to build model (sec)	0.84	0.14	0.04	0.02
Accuracy (%)	97.89	93.14	92.61	91.73

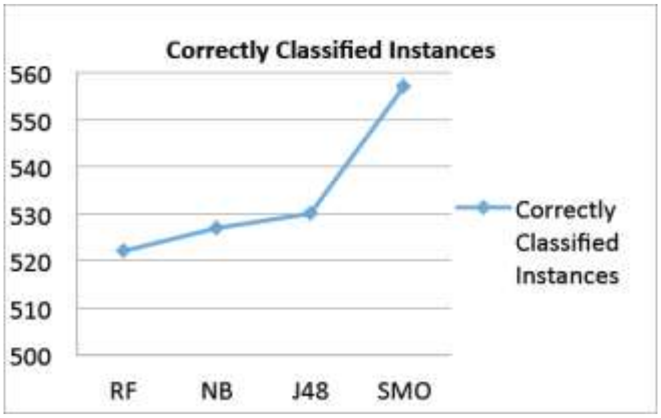


Fig.:1 Comparison of classifiers based on Correctly Classified Instances.

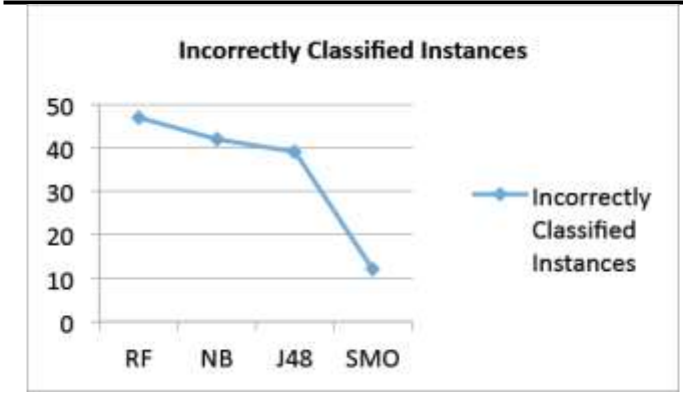


Fig.:2 Graph illustrating the comparison of classifiers based on their Incorrectly Classified Instances

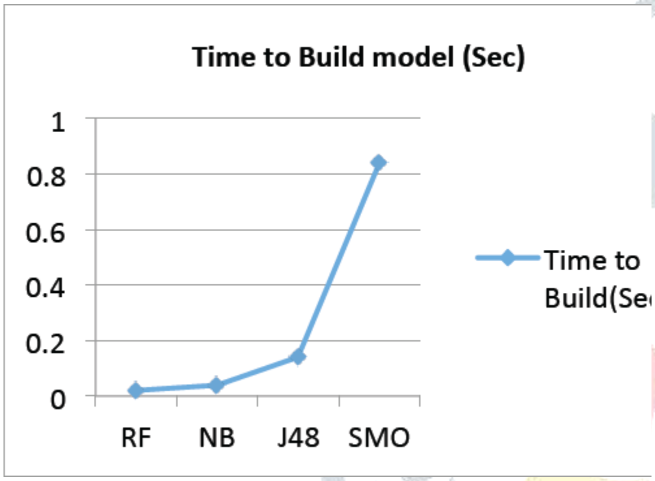


Fig.:3 Graph depicting the comparison of classifiers regarding the time taken to build a model

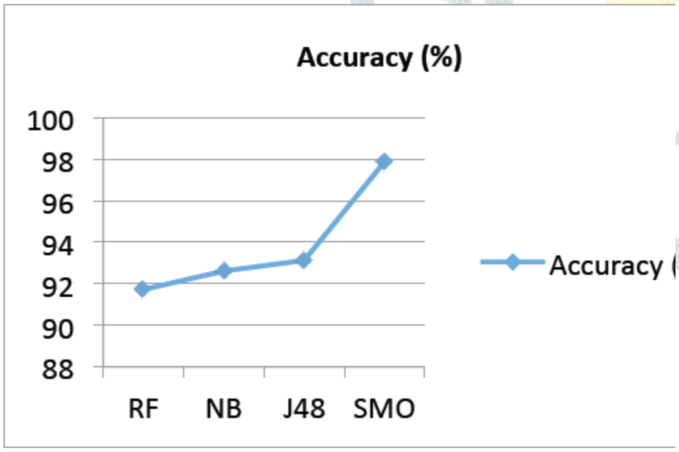
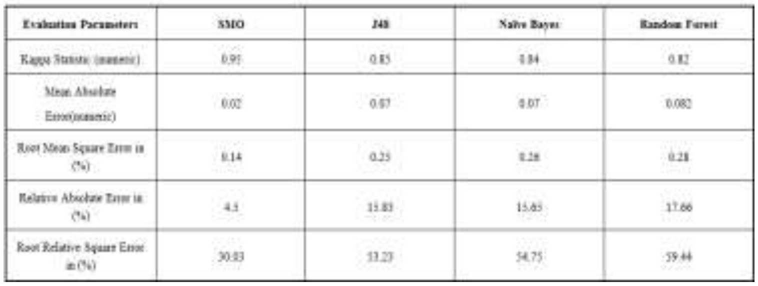


Fig.:4 Graph comparing classifiers based on their accuracy.

Table 3. Training and Simulation Output.



Evaluation Parameters	SMO	J48	Naïve Bayes	Random Forest
Kappa Statistic (numeric)	0.95	0.85	0.84	0.82
Mean Absolute Error(numeric)	0.02	0.07	0.07	0.082
Root Mean Square Error in (%)	0.14	0.25	0.26	0.28
Relative Absolute Error in (%)	4.5	15.83	15.65	17.66
Root Relative Square Error in (%)	30.03	53.23	54.75	59.44

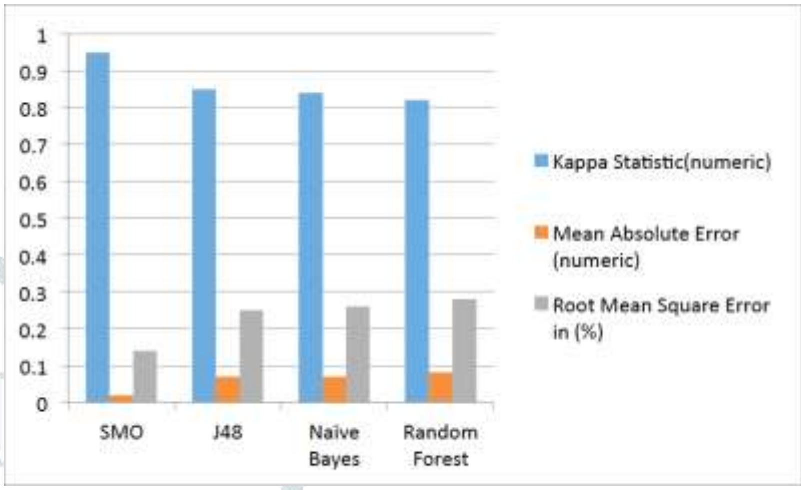


Fig.:5 Graph illustrating the comparison of classifiers based on Kappa Statistic, Mean Absolute Erro

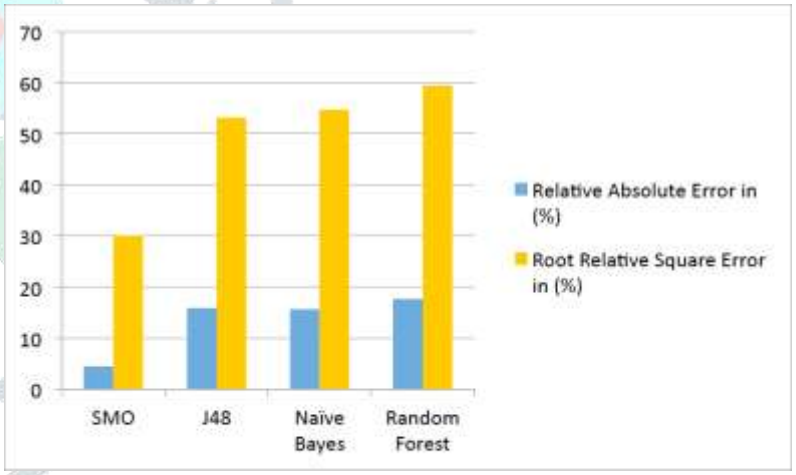
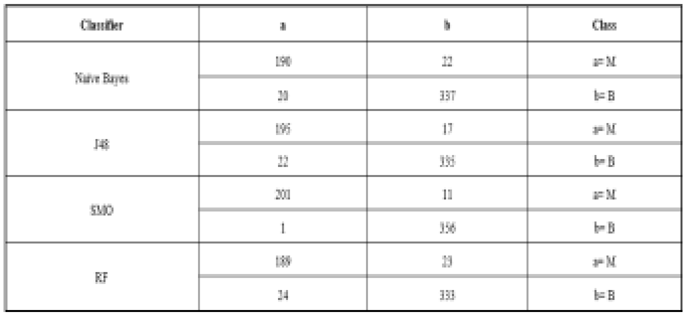


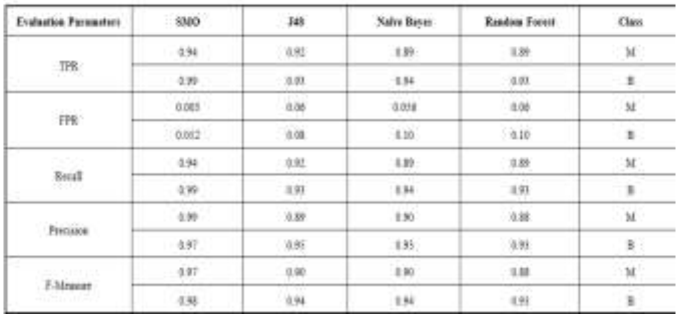
Fig.:6 Graph depicting the comparison of classifiers based on Relative Absolute Error and Root Rel

Table 4. Confusion Matrix of the classifiers.



Classifier	a	b	Class
Naïve Bayes	190	22	a= M
	20	337	b= B
J48	195	17	a= M
	22	335	b= B
SMO	201	11	a= M
	1	356	b= B
RF	189	23	a= M
	24	333	b= B

Table 5. Comparative analysis of classifiers to check the efficiency.



Evaluation Parameters	SMO	J48	Naïve Bayes	Random Forest	Class
TPR	0.94	0.92	0.89	0.89	M
	0.99	0.93	0.94	0.93	B
FPR	0.003	0.06	0.038	0.06	M
	0.012	0.08	0.10	0.10	B
Recall	0.94	0.92	0.89	0.88	M
	0.99	0.93	0.94	0.93	B
Precision	0.99	0.89	0.90	0.88	M
	0.97	0.95	0.95	0.93	B
F-Measure	0.97	0.90	0.90	0.88	M
	0.98	0.94	0.94	0.93	B

In this work, training and simulation errors are also considered for a more thorough evaluation of the classifiers' performance, as shown in Table 3 and Figures 5/6. By comparison with other classifiers, these show that SMO has the lowest error rate (0.14%).

The categorization models used in our investigation are represented by the confusion matrix in Table 4. It provides an overview of how a particular classifier is expected to perform. The projected classes are shown by the rows in this table, while the actual classes are indicated by the columns. Class A represents cases of malignant breast cancer, and Class B represents cases of benign individuals with breast cancer. Based on these two classes, predictions are produced.

Using parameters from the confusion matrix, the classifiers are assessed.

Sensitivity== TP/ (TP+FN)………………………..(3)

Specificity == TN/ (FP+TN)………………………(4)

Recall ==TP/ (TP+FN)……………………………(5)

Precision == TP/ (TP+FN)……………………….(6)

F-Measure == 2(P*R)/P+R……………………….(7)

Where
P= =Precision,
R==Recall,
F== False

CONCLUSION

Four classification algorithms were used in this Nobel research and study to forecast results on the Wisconsin Breast Cancer dataset (WDBC): SMO, J48, Naïve Bayes, and Random Forest (RF). To select the best classifier for the dataset, we evaluated each classifier's efficacy and efficiency using a variety of criteria. Comparing SMO to the other classifiers used in our study to predict breast cancer as benign or malignant, SMO produced the greatest accuracy rate of 97.89% with a low error rate of 0.14% in WEKA. SMO was used to train Support Vector Machine (SVM). These results imply that classification algorithms can be useful instruments in the medical domain for the diagnosis of long-term illnesses like cancer.

PREFERENCES

- "Accurate brain tumor detection using deep convolutional neural network", Elsevier, Computational and Structural Biotechnology Journal 20 (2022) 4733–4745, Md. Saikat Islam Khan, Anuchur Rehman Sarwar, Tanoy Debnath, Md. Razual Karim, Mostofa kamal Nasir, Shahab S. Band, Amir Mosawi, and Iman Dehzangi.
- The article "ICT in Healthcare" was written by A. Sarwar, J. Manhas, and V. Sharma and was published in The Stances of –Government Policies, Processes, and Technologies, 2018, pp. 31–40.
-] Ramachandran, P., Girija, P.N., and Bhuvaneswari, T. (2014) "Data mining approaches for early cancer detection and prevention," International Journal of Computer Applications, vol. 97, no. 13, pp. 48–53.
- In the International Journal of Computer Applications, Ramachandran, P., Girija, P.N., and Bhuvaneswari, T. (2014), "Data mining techniques for early cancer diagnosis and prevention," vol. 97, no. 13, pp. 48–53.
-] "A breast cancer decision support system for rural people," by Thoranin Intarajak and Seung Hwan Kang, International Journal of Computer, the Internet, and Management, vol. 17, no. SP1, pp. 47.1–47.8, March 2009.
- In the Indian Journal of Fundamental and Applied Life Sciences, vol. 5, no. S1, pp. 4330–4339, 2015, H. Karim and K. Zand published "A comparative assessment on data mining strategies for breast cancer detection and prediction."
- Kehinde Williams, Peter Adebayo Idowu, Jeremiah Ademola Ba- logun, and Adeniran Ishola Oluwaranti, "Breast cancer risk predic- tion using data mining classification algorithms", Transactions on Networks and Communications, vol. 3, no. 2, pp. 01-11, May 2015. [http://dx.doi.org/10.14738/tnc.32.662]
- A. Ahmadi and P. Afshar, "Using particle swarm optimization and support vector machines for intelligent breast cancer identification," Journal of Experimental Theory and Artificial Intelligence, vol. 28, no. 6, pp. 1021-1034, 2016. [10.1080/0952813X.2015.1055828] (http://dx.doi.org)
- V. Chaurasia, S. Pal, and BB Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques," Journal of Algorithms and Computer Technology, vol. 12, no. 2, pp. 119–126, 2018. [10.1177/1748301818756225] can be found online.
- Tran T. and Le U., "A Data Mining Approach for Predicting Breast Cancer Risk,"

International Conference on the Development of Bio-medical Engineering, 2017, pp. 223-228

- An Effective Prediction of Breast Cancer Data Using Data Mining Techniques, G. Ravi Kumar, G.A. Ramachandra, and K. Nagamani, International Journal of Innovations in Engineering and Technology, vol. 2, no. 4, pp. 139–144, 2013.
- In the 6th International Symposium on Frontiers in Ambi-ent and Mobile Systems Procedia Computer Science, vol. 83, 2016pp. 1064-1069, Asria, H., Mousannifb, H. Al Moatassimec, and T. Noeld, "Us- ing machine learning algorithms for breast cancer risk prediction and diagnosis" [http://dx.doi.org/10.1016/j.procs.2016.04.224]
- The article "Clustering algorithm for a healthcare dataset using silhouette score value" was published in the International Journal of Computer Science & Information Technology in 2018. It was authored by G. Ogbuabor and F.N. Ugwoke. [10.5121/ijcsit.2018.10203] (http://dx.doi.org)
- "Data mining in healthcare: a review" by Nur' Aini Abdul Rashid and Wahidah Husain, presented at the Third Information System International Conference, by N. Jothi, Procedia Comput. Sci., vol. 72, pp. 306-313, 2015. [www.doi.org/10.1016/j.procs.2015.12.145]
- Data Mining and Knowledge Engineering, vol. 6, no. 1, pp. 21-29, 2014; Saleema, J.S., P. Deepa Shenoy, K.R. Venugopal, and M. Lalith, Patnaik. "Cancer prognostic prediction model using data mining approaches."
- Data mining: principles, models, techniques, and algorithms, by M. Kantardzic, John Wiley & Sons, 2003.
- Springer: New York, 2005. Data Mining and Knowledge Discovery Handbook, vol. 2.
- X. Wu and J. Vipin Kumar "Top 10 algorithms in data mining: Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McClachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg [http://dx.doi.org/10.1007/s10115-007-0114-2]
- Flags Data Set, UCI Machine Learning Repository. [Referenced April 18, 2019]. Accessible at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wiscon in+%28Diagnostic%29
- Applying data mining for the study of breast cancer data is a work by D.M. Liou and W.P. Chang [http://dx.doi.org/10.1007/978-1-4939-1985-7_12].