Journal of Computer Science

Original Research Paper

Classify Breast Cancer Patients using Hybrid Data-Mining Techniques

[1]Faris E. Mohammed, [2]Nadia Smaoui Zghal,
[3]Dalinda Ben Aissa and [4,5]Mostafa Mahmoud El-Gayar

[1]Microwave Electronics Research, Tunis University, Tunisia
[2]Control and Energy Management Research Laboratory, ENIS, Sfax, Tunisia., ENIS, Sfax, Tunisia, Tunisia
[3]Microwave Electronics Research Laboratory, Faculty of Science, Tunis University, Liberia
[4]Department of information technology, Faculty of Computers, and Information, Mansoura University, Mansoura, Egypt
[5]Faculty of Computer science and engineering, New Mansoura University, Gamsa, Egypt

Article history
Received: 20-03-2022
Revised: 14-04-2022
Accepted: 19-04-2022

Corresponding Author:
Faris E. Mohammed
Microwave Electronics Research, Tunis University, Tunisia
Email: tchangmsing@gmail.com

Abstract: According to the World Health Organization (WHO), breast cancer is a disease that leads to death, especially for women who have neglected or ignored the risk factors. Doctors can classify patients according to clinical information, famous disease symptoms, or similar cases. But, some cases are difficult to detect early or diagnose accurately. Therefore, the most important challenge faced by researchers in this field is how to classify patient data by extracting important information that leads to the detection of the disease early and correctly. This article proposes the enhanced system of a decision support system based on hybrid classification algorithms to classify Breast cancer patients accurately and quickly. The main contribution of this article is to develop an algorithm that filters the data and solves the problem of missing data in some records to facilitate the classification of data. In the experiments conducted, the proposed system was learned by several algorithms on a standard Electronic Health Records (HER) dataset to determine the appropriate test factors. Four experiments were performed to measure the accuracy and speed of the different data mining techniques. The proposed ensemble process achieved a high accuracy rate up to 99% in a good time.

Keywords: Breast Cancer, Datamining, Electronic Health Records, Decision Tree, Random Forest

Introduction

Background

More than 20 million patients have died around the world, according to (WHO) report for 2016 due to cancer, especially breast cancer. Breast cancer is the second most common cancer in women compared to all other cancers. The report said about 4 million of these deaths occurred in young people's lives (Kathale and Thorat, 2020) however, despite large numbers of statistics, many stories from WHO and several hospitals say that 90% of these cases can be saved and prevented if patients are diagnosed correctly and early. Hospitals are traditional environments for health care (Jain and Srivastava, 2013). These environments depend on the diagnosis of diseases on medical equipment and doctors' reports to determine the symptoms and causes of various diseases. Any unintentional error may result from overcrowding or lack of resources. So, an alternative solution can be used that reduces the error rate and works faster by applying a high-speed expert system (Patra *et al*., 2019; Laghmati *et al*., 2019; Prasetyo *et al*., 2014). There are various important challenges in this field which can be summed up as follows:

- The presence of missing information in the patient's data may lead to a misdiagnosis
- Late diagnosis of breast cancer may lead to patient death
- Extracting and analyzing essential information for a large group of patients may take much time and lead to errors

The idea of a decision support system is to address the limitations and problems associated with traditional hospital diagnoses. The process of discovering knowledge from extensive data consists of many significant and systematic steps, as presented in Fig. 1.

Science Publications

© 2022 Faris E. Mohammed, Nadia Smaoui Zghal, Dalinda Ben Aissa, and Mostafa Mahmoud El-Gayar. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

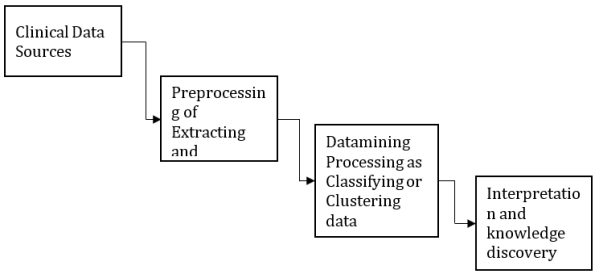

Fig. 1: Steps of discovering knowledge from clinical sources Laghmati *et al*. (2019)

Many data mining and artificial intelligence approaches are used in the clinical and healthcare field (Omran *et al*., 2021; Ashri *et al*., 2021; Singh *et al*., 2018; Liu *et al*., 2009). Data mining techniques are divided into some categories such as classification, association, and clustering.

Ilhan *et al*. (2020) identified a deep learning technique that supplies a potent tool to assist experts to analyze, modeling, and make sense of complex clinical data in a wide range of medical applications. The goal of this study is to develop an effective system for classifying breast tumors as benign and malignant. The accuracy of the developed approach is 98.42%.

Lahoura *et al*. (2021) suggested a cloud-based ELM with applied machine learning algorithms on the Wisconsin Diagnostic Breast Cancer (WBCD) dataset. ELM's best performance was achieved in both standalone and cloud environments and a comparison has been made. The conclusions of the experimental results indicate that the accuracy obtained is 0.9868, the precision 0.9054, the recall 0.9130, and the F1 score 0.8129.

Ara *et al*. (2021) developed an automatic classifier system using multi-classifiers for breast cancer diagnosis. The precision of each method was calculated and compared to find the most appropriate one. Random Forest and SVM achieved higher accuracy of 96.5%.

Sengar *et al*. (2020) suggested a proposed system that uses Logistic Regression and Decision Tree algorithms on the Wisconsin (Diagnostic) Data Set. The results have shown that the proposed system has a 90% accuracy rate.

Chaurasia and Pal (2017) match the implementation standard for supervised learning classifiers, such as SVM and Decision Tree (J48) with a classification and regression tree (CART) to discover the most appropriate classifier in breast cancer datasets. The practical result indicates that the J48 kernel and SVM are more accurate than the other methods. achieves an accuracy of 96.84% in the Wisconsin breast cancer datasets.

Dinesh *et al*. (2018) proposed a system for classifying cardiac patients using the Naïve Bayes algorithm. The dataset used is being received by one among such leading academy to study diabetes in Chennai. In their experiments, they used the WEKA instrument with a proportional seventy split to perform the classification method. The results show that Naive Bayes has an accuracy rate of 86.419%.

According to these related works, we found some limitations such as (Habib *et al*., 2020; Abbas *et al*., 2021):

- The presence of missing information in the patient's data may lead to a misdiagnosis.
- Overfitting problems.
- Extracting and analyzing many features for a large group of patients may take much time and lead to errors (low accuracy).

Proposed Methods

In this section, the block diagram of the proposed system is shown in Fig. 2. The proposed system passed through three stages (preprocessing stage, classification stage, and decision stage), and each stage will be described in the next subsections. Each blue box represents one of the main contributions of this article.

Data Preprocessing Stage

This stage consists of four blocks. The first block is responsible for uploading the original dataset into the system. The second block is one of the main contributions to this article and consists of three sub-steps.

The first sub-step parses the input file and initializes schema based on the type of input data. The second sub-step is filtering the original data to remove duplications, perform stemming and extract records that have missing values.

The third block is calculating the average, variance, and standard deviation for each column. For example, if we have an age column with some numeric values, then we can calculate the average and standard deviation for that column. Finally, the fourth block is used to replace each missing value with the Maximum Likelihood Estimation (MLE) value from equation (1). Then the new MLE value is added to the (JavaScript Object Notation) JSON file. After that, the JSON file is converted to the database for further use.

$$P(x;\mu,\sigma)=\frac{1}{2\pi\sigma^2}e\frac{-\left(x-\mu^2\right)}{2\sigma^2} \quad (1)$$

In the proposed schema type identification method described in algorithm 1, the originally uploaded dataset is considered as input, and the initialized schema of JSON format and data type for each column is considered as output.

Algorithm#1 Schema type identification

Input ← Dataset (**DS**)
Output ← Data type and schema
1. Start Method
2. connect ← connection. Open (**DS**)
3. length ← connect. Count (**DS**. records)

```
4.      Schema ← null
5.      Data_type ← null
6.      For i ← 0 to length do
7.        while row ∈ DS. set(i) do
8.          while column_name ∈ row do
9.            Schema ← schema ∪ column_name
10.           If row.typeformat == int
11.                Data_type = int
12.             Else If record.scanformat == float
13.               Data_type = float
14.           Else
15.              Data_type = string
16.           End IF
17.       End while
28.     End while
29.   End For
20.   Get Schema and data type
21. End Schema Type Identification method
```

In the proposed parsing and filtering method described in algorithm 2, the initially uploaded dataset, schema of the previous step, the row of data, and data type for each column are considered as input and the final JSON file format with the row data is considered as output, as shown in Fig. 3. This algorithm uses three main processes to transform data, such as data cleaning, duplication removal, and binding data into JSON format.

Algorithm#2 Parsing and Filtering Method

```
Input ← DS, Row Data, Schema, Data type
Output ← Data in JSON format
1. Start Method
2. Initialize schema (Schema)
3. Limit ← DS. Row Data. Length
4. Initial_number ← 1
5. While initial < Limit do
6. Clean (Row Data) // skip records with
7. missing values
8. Remove Duplicates (Row Data)
9. Schema ← encodes Row Data using
10. JSON format (Data type)
11. Initial_ number ++
12. End While
13. Return JSON
14. End Parsing and Filtering Method
```

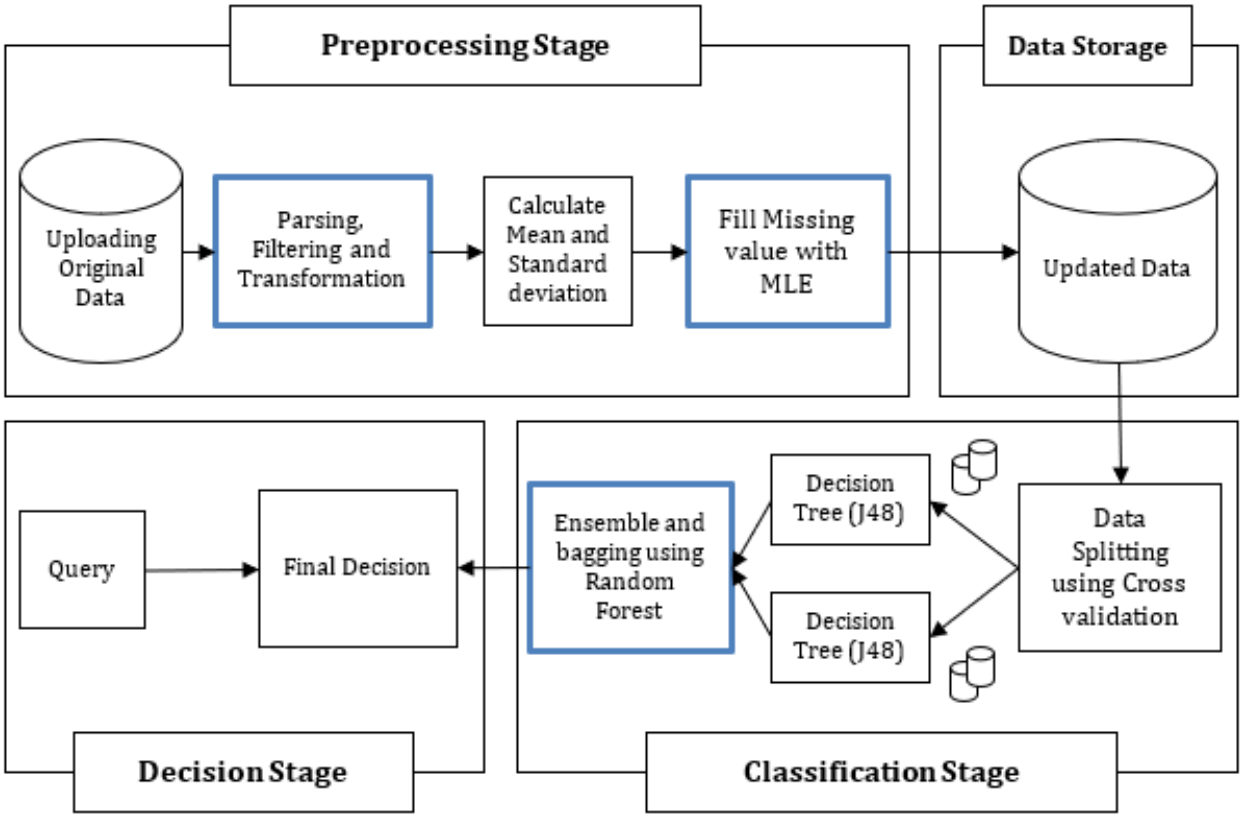

Fig. 2: The proposed system

```
Patient_id
{
  gender: value (numeric or string)
  age: value (numeric)
  history: value (numeric or string)
  breast_features
  {
   Texture
   {
      mean: value (float)
      Standard_deviation: value (float)
   }
   Radius
   {
      mean: value (float)
      Standard_deviation: value (float)
   }
   Smoothness
   {
      mean: value (float)
      Standard_deviation: value (float)
   }
   area
   {
      mean: value (float)
      Standard_deviation: value (float)
   }
  }
```

Fig. 3: Sample of the JSON structure

Classification Stage

This stage consists of three main blocks. The first block is responsible for splitting the processed dataset into various samples for the training system on datasets using the cross-validation technique. Ten-fold cross-validation was used. Ten-fold cross-validation divides the data set into 10 parts. Nine parts are then used for training and the 10th part is used for testing. The second block is the main contribution in this article and consists of two sub-steps. The first sub-step is applying implemented decision tree (J48) algorithm using equation (2) to train the system in the dataset.

$$E = \sum_{i=1}^{k} P_i \log_2 P_i \tag{2}$$

where k is that value from classes of this objective feature. P_i is a value of incidences from class i separated via the whole value from occurrences. The decision tree algorithm creates rules that are used in the final decision. A sample of these rules is shown in Fig. 4. Then in the second sub-step, the ensemble process is developed using the Random Forest algorithm. Bootstrap aggregation is a universal process that can be applied to decrease the adjustment for those methods that have high adjustment. Random Forests are an enhancement over bagged decision trees.

```
area_ws <= 880.8
|   concave points_ws <= 0.1357: 0 (337.0/7.0)
|   concave points_ws > 0.1357
|   |   texture_ws <= 27.37
|   |   |   concave points_ws <= 0.1789
|   |   |   |   area_se <= 21.91: 0 (12.0)
|   |   |   |   area_se > 21.91
|   |   |   |   |   perimeter_se <= 2.615: 1 (6.0/1.0)
|   |   |   |   |   perimeter_se > 2.615: 0 (6.0)
|   |   |   concave points_ws > 0.1789: 1 (4.0)
|   |   texture_ws > 27.37: 1 (21.0)
area_ws > 880.8
|   concavity_mean <= 0.0716
|   |   texture_mean <= 19.54: 0 (9.0/1.0)
|   |   texture_mean > 19.54: 1 (10.0)
|   concavity_mean > 0.0716: 1 (164.0)
```

Fig. 4: Sample of rules from the decision tree

Experimental Results

System Specification

The proposed system is evaluated using one machine and a local server. The machine has an Intel Core i-5 processor@2.2GHz with 4GB RAM. The local server is an Apache server with a local database management engine called MYSQL. The program tool used to perform data mining tasks is WEKA version 3.6.9.

Datasets

The Breast Cancer dataset used in this analysis is provided by the UCI storage located in Wisconsin. The dataset statistical information was represented in Table 1. The dataset has 569 records with two classes (malignant and benign) and nine numeric features, as described in Table 2.

Performance Metrics

We use a 10-fold cross-validation technique to avoid overfitting problems. The accuracy of the applied procedures must be evaluated applying under titles about, False Positive (FP) rate, True Positive (TP) rate, False Negative (FN) rate, True Negative (TN) rate, recall, precision, ROC area, CPU Time, accuracy, error. We used Eq. (3,4 and 5) to evaluate the methods. These equations are listed below:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$Accuracy = \frac{2\,precision * recall}{precision + recall} \times 100\% \tag{5}$$

Performance Evaluations

In this sub-section, we applied four experiments using different data mining algorithms on the original dataset

after performing preprocessing techniques. We found that the neural network and random forest achieved high accuracy. But, the neural network algorithm has taken much time to complete the classification process. Also, we found that the decision tree algorithm (J48) and Naïve Bayes method achieved less performance than other methods. But, they had taken minimum time to perform the classification process. So, we need to perform an ensemble process (hybrid) between the method with the highest performance and with fastest one to achieve the best results.

Performance Results of Naïve Bayes Algorithm

In Table 3, the performance results as precision, recall, and accuracy are 93, 93, and 93% respectively. It has applied in 0.07 sec.

Performance results of Decision Tree Algorithm

In Table 4, the performance results as precision, recall, and accuracy are 95, 94 and 94% respectively. It is applied in 0.17 sec.

Performance Results of Artificial Neural Network Algorithm

In Table 5, the performance results in precision, recall, and accuracy are 97, 95, and 96% respectively. It has applied in 7.39 sec (very slow).

Performance Results of Random Forest Algorithm

In Table 6, the performance results as precision, recall, and accuracy are 96, 95, and 95% respectively. It has applied in 0.59 sec.

Performance measurements of Hybrid Algorithms

In Table 7, we discuss the results of the proposed hybrid process (ensemble process) between decision tree and random forest algorithms. We applied 10-fold cross-validation for splitting records into samples. Decision tree algorithm was applied on each sample with different number of iterations (k = 25, K = 50, K = 75, K = 100 and k = 125). Then all samples were bagged and ensembled using the random forest technique for a final decision. We found that the best performance was achieved at k = 100. We noticed that the performance results in precision, recall, and accuracy are 98, 96, and 97% respectively. It has applied in 0.19 seconds. Also, we found that the proposed ensemble process achieved higher performance than other methods in a good time as shown in Fig. 5 and 6.

Table 1: Dataset statistical information

Patient Records	569 Records
Attributes	35
Gender	469 Female
	100 Male
Range of ages	10 – 90 years

Table 2: Features description

Features	Description
Radius	mean of distances from center
Texture	Standard deviation of gray-scale
Area	Numeric values
Smoothness	Local variation in radius lengths
Compactness	Perimeter^2/area-1
Concavity	Severity of concave portions
Concave Points	Number of concave portions
Fractal Dimension	Coastline approximation -1

Table 3: Experimental Result of Naïve Bayes Algorithm.

Metric	Percentages
Precision	93%
Recall	93%
Accuracy	93%
Time	0.07 sec

Table 4: Experimental result of naïve Bayes algorithm

Metric	Percentages
Precision	95%
Recall	94%
Accuracy	94%
Time	0.17 sec

Table 5: Experimental result of Artificial Neural Network (ANN) algorithm

Metric	Percentages
Precision	97%
Recall	95%
Accuracy	96%
Time	7.39 sec

Table 6: Experimental results of random forest algorithm

Metric	Percentages
Precision	96%
Recall	95%
Accuracy	95%
Time	0.59 sec

Table 7: Experimental results of proposed ensemble process

Iteration	Metric	Values
K = 25	Precision	92%
	Recall	92%
	Accuracy	92%
	Time	0.11 sec
K = 50	Precision	94%
	Recall	92%
	Accuracy	93%
	Time	0.15 sec
K = 75	Precision	96%
	Recall	94%
	Accuracy	95%
	Time	0.17 sec
K = 100	Precision	99%
	Recall	99%
	Accuracy	99%
	Time	0.19 sec
K = 125	Precision	97%
	Recall	95%
	Accuracy	96%
	Time	0.21 sec

Statistical Measures

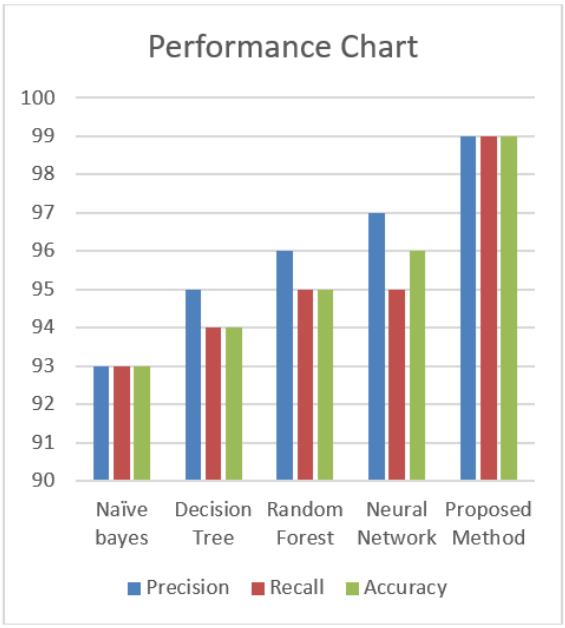

Fig. 5: Performance chart between different data mining algorithms and proposed ensemble method

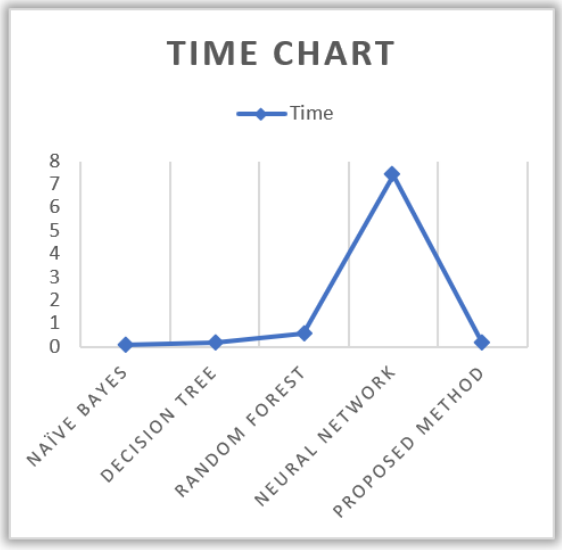

Fig. 6: Time chart between different data mining algorithms and proposed ensemble method

Conclusion

Currently, data mining algorithms play an important role in diagnosing and classifying patient data. In this study, a proposed system for the classification and prediction of breast cancer is presented. This proposed system is divided into several parts, including the initial processing, in which the data is initially processed before entering the data mining algorithms, and then comes the role of data classification. A hybrid model was used between two algorithms that achieved the highest accuracy rate. The proposed system analyzes patient data and categorizes it into groups for training and testing to identify, analyze and predict mixed disorders and syndromes. Finally, the empirical results showed that the proposed hybrid classifier offers the most favorable achievement in terms of accuracy through the large data set.

Future Work

In the future, we can use AI techniques and genetic algorithms to select the most suitable features to minimize the consumption of time such as recommended algorithm used in (Ashri *et al*., 2021). Also, we need to analyze image datasets using deep learning that will help for faster and better detection of the different types of breast cancer.

Authors Contributions

Faris E. Mohammed: Writing of the manuscript.
Nadia Smaoui Zghal: Nadia Smaoui Zghal.
Dalinda Ben Aissa: Organized the study.
Mostafa Mahmoud El-Gayar: Participated in all experiments and coordinated the data-analysis

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abbas, S., Jalil, Z., Javed, A. R., Batool, I., Khan, M. Z., Noorwali, A., ... & Akbar, A. (2021). BCD-WERT: a novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm. PeerJ Computer Science, 7, e390. https://peerj.com/articles/cs-390/

Ara, S., Das, A., & Dey, A. (2021, April). Malignant and benign breast cancer classification using machine learning algorithms. In 2021 International Conference on Artificial Intelligence (ICAI) (pp. 97-101). IEEE. doi.org/10.1109/ICAI52203.2021.9445249

Ashri, S. E., El-Gayar, M. M., & El-Daydamony, E. M. (2021). HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm. IEEE Access, 9, 146797-146809. doi.org/10.1109/ACCESS.2021.3122789

Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. International journal of innovative research in computer and communication engineering (An ISO 3297: 2007 Certified Organization) Vol, 2.

Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-7). IEEE. doi.org/10.1109/ICCTCT.2018.8550857

Habib, P. T., Alsamman, A. M., Hassnein, S. E., Shereif, G. A., & Hamwieh, A. (2020). Assessment of Machine Learning Algorithms for Prediction of Breast Cancer Malignancy Based on Mammogram Numeric Data. medRxiv. https://www.medrxiv.org/content/10.1101/2020.01.08.20016949v1

Ilhan, U., Uyar, K., & Iseri, E. I. (2020). "Breast Cancer Classification Using Deep Learning," pp. 709–714, Aug. 2020. doi.org/10.1007/978-3-030-64058-3_88

Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. IJRET: International Journal of Research in Engineering and Technology, 2(11), 2319-1163. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.802&rep=rep1&type=pdf

Kathale, P., & Thorat, S. (2020, February). Breast cancer detection and classification. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE.doi.org/10.1109/ic-ETITE47903.2020.367

Laghmati, S., Tmiri, A., & Cherradi, B. (2019, October). Machine learning based system for prediction of breast cancer severity. In 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 1-5). IEEE. doi.org/10.1109/WINCOM47513.2019.8942575

Lahoura, V., Singh, H., Aggarwal, A., Sharma, B., Mohammed, M. A., Damaševičius, R., ... & Cengiz, K. (2021). Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. Diagnostics, 11(2), 241. doi.org/10.3390/diagnostics11020241

Liu, Y. Q., Wang, C., & Zhang, L. (2009, June). Decision tree based predictive models for breast cancer survivability on imbalanced data. In 2009 3rd international conference on bioinformatics and biomedical engineering (pp. 1-4). IEEE. doi.org/10.1109/ICBBE.2009.5162571

Omran, N. F., Abd-el Ghany, S. F., Saleh, H., & Nabil, A. (2021). Breast cancer identification from patients' tweet streaming using machine learning solution on spark. Complexity, 2021. doi.org/10.1155/2021/6653508

Patra, R., & Khuntia, B. (2019, February). Predictive analysis of rapid spread of heart disease with data mining. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-4). IEEE. doi.org/10.1109/ICECCT.2019.8869194

Prasetyo, C., Kardiana, A., & Yuliwulandari, R. (2014). "Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques," Int. J. Adv. Res. Artif. Intell., 3, 2014. doi.org/10.14569/IJARAI.2014.030703

Sengar, P. P., Gaikwad, M. J., & Nagdive, A. S. (2020, August). Comparative study of machine learning algorithms for breast cancer prediction. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 796-801). IEEE. doi.org/10.1109/ICSSIT48917.2020.9214267

Singh, A. S., Irfan, M., & Chowdhury, A. (2018, December). Prediction of liver disease using classification algorithms. In 2018 4th international conference on computing communication and automation (ICCCA) (pp. 1-3). IEEE. doi.org/10.1109/CCAA.2018.8777655