Asian Journal of Research in Computer Science

8(4): 45-59, 2021; Article no.AJRCOS.68450
ISSN: 2581-8260

A Comparative Analysis and Predicting for Breast Cancer Detection Based on Data Mining Models

Shler Farhad Khorshid[1*], Adnan Mohsin Abdulazeez[2] and Amira Bibo Sallow[3]

[1]Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.
[2]Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.
[3]Nawroz University, Duhok, Kurdistan Region, Iraq.

Authors' contributions

This work was carried out in collaboration among all authors. Author SFK managed the literature searches related to breast cancer classification and wrote the first draft of the manuscript. Author AMA gave the idea and designed the study. Author ABS performed the statistical analysis data and discuss the results. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v8i430209
Editor(s):
(1) Dr. G. Sudheer, GVP College of Engineering for Women, India.
Reviewers:
(1) S. Rajasekaran, University of Technology and Applied Sciences-Ibri, Oman.
(2) D. Mallikarjuna Reddy, VIT University, India.
(3) Tesfay Gidey Hailu, Addis Ababa Science and Technology University, Ethiopia.
Complete Peer review History: http://www.sdiarticle4.com/review-history/68450

Review Article

Received 10 March 2021
Accepted 14 May 2021
Published 19 May 2021

ABSTRACT

Breast cancer is one of the most common diseases among women, accounting for many deaths each year. Even though cancer can be treated and cured in its early stages, many patients are diagnosed at a late stage. Data mining is the method of finding or extracting information from massive databases or datasets, and it is a field of computer science with a lot of potentials. It covers a wide range of areas, one of which is classification. Classification may also be accomplished using a variety of methods or algorithms. With the aid of MATLAB, five classification algorithms were compared. This paper presents a performance comparison among the classifiers: Support Vector Machine (SVM), Logistics Regression (LR), K-Nearest Neighbors (K-NN), Weighted K-Nearest Neighbors (Weighted K-NN), and Gaussian Naïve Bayes (Gaussian NB). The data set was taken from UCI Machine learning Repository. The main objective of this study is to classify breast cancer women using the application of machine learning algorithms based on their accuracy. The results have revealed that Weighted K-NN (96.7%) has the highest accuracy among all the classifiers.

**Corresponding author: E-mail: shler.sulayvani@gmail.com;*

Keywords: Breast cancer; data mining; SVM; Logistics regression; weighted K-NN; Gaussian Naïve Bayes.

1. INTRODUCTION

Data mining (DM) uses a variety of techniques (such as classification, clustering, regression, association rules, and so on) and algorithms (such as Decision Tree (DT), Genetic Algorithms, Nearest Neighbor Form, and so on) to analyze large amounts of raw or multi-dimensional data. To put it another way, DM can derive hidden knowledge from large databases of clinical or medical data obtained from health centers or hospitals using intelligent data analysis. These insights can help enhance decision-making, prevention, diagnosis, and treatment in the field of medicine [1], [2], [3], [4], [5]. Furthermore, DM may establish relationships or define association rules between various features, such as a patient's personal details, disease symptoms, and so on [6], [7].

In the field of medicine, DM plays a significant role in computing applications [8], [9], [10]. The applications and methods of DM are demonstrated in the areas of healthcare administrations, patient care, management, and intensive care systems. Breast Cancer (BC) is the most common of all cancers, and it is the leading cause of cancer deaths in women around the world, according to one of the latest DM studies [11], [12]. BC is one of the diseases with the highest number of cases and deaths worldwide [13], [14], [15]. After lung cancer, it is the second leading cause of death in women [16]. There are two types of breast tumors: malignant and benign [17], [18], [19]. As cells in the breast tissue become isolated and without the usual controls on cell passing and cell division, a malignant tumor develops [20]. Benign tumors have a good contour and are not harmful. They grow slowly in the organ where they first appeared, with no signs of metastatic disease [21]. Benign tumors are made up of cells that look like normal breast tissue cells. While malignant ones are harmful because they can expand to other parts of the body and cause metastatic disease. Cancer cells in malignant tumors have many abnormalities in the form, scale, and contour as compared to normal cells, where cells lose their original characteristics. DM algorithms can be a useful tool for predicting and diagnosing breast cancer, as well as classifying them into; benign or malignant tumors. Earlier treatment of BC can result in curing the body of this disease [22].

In this paper, a comparative analysis is presented of five different DM classification algorithms namely LR, SVM, K-NN, Weighted K-NN, and Gaussian NB on the Breast Cancer Data Set by measuring their classification accuracy. Results show that all the presented DM algorithms performed well on the classification task.

The rest of the paper is organized as follows. Section 2 is a presentation of breast cancer. Section 3 discusses data mining. Section 4 focused on some of the applications of DM. Section 5 gives a review of similar research. The material and methods used in the working process are discussed in section 6. Section 7 summarizes the findings and discusses them. Section 8 provides a comparison of the related works. Finally, Section 9 presents the conclusion of this paper.

2. BREAST CANCER (BC)

BC grows as cells in the breast tissue differentiate and expand without the usual controls on cell division and death [23]. It's the most common form of cancer in women [24], [25]. While experts do not know the precise causes of the majority of breast cancers, they do know some of the risk factors that increase a woman's chances of contracting the disease. Age, genetic risk, and family history are examples of these influences [26].

BC treatments are divided into two categories, regional and systematic. Systematic treatments include chemotherapy and hormone therapy, while regional treatments include surgery and radiation. The two forms of the treatment are frequently utilized together to obtain better results. Despite the fact that BC is the second most common cause of death in women, it has a high survival rate. Ninety- seven percent of women live for five years or longer if they are diagnosed at an early stage [27].

3. DATA MINING (DM)

In the domain of medicine, DM is playing a major role in computing applications [28], [29]. DM research relies on classification algorithms. Classification, clustering, association rules, prediction, and neural networks are some of the DM applications and techniques that are used to

analyze large amounts of data. Among these, some classification algorithms such as Naïve Bayes (NB), SVM, Artificial Neural Network (ANN), DT (C 5.0) and K-NN algorithms to are utilized to achieve the most accurate results. DM is currently being used to solve a variety of real-world issues since the primary aim of DM is to convert raw data into more usable information. Medical databases raise problems for pattern extortion due to their complex features [30], [31], [32], [33]. DM algorithms can be classified into two types: statistical and machine learning (ML) algorithms. DM processes are categorized into descriptive and predictive categories (Fig.1). Descriptive mining tasks show the database's general data properties. To perform predictive mining tasks, the inference is made on the results whereby a forecast is rendered based on explicit values defined by established results. Descriptive data mining offers characteristics and definitions for the data set without the need for a predefined objective [34], [35].

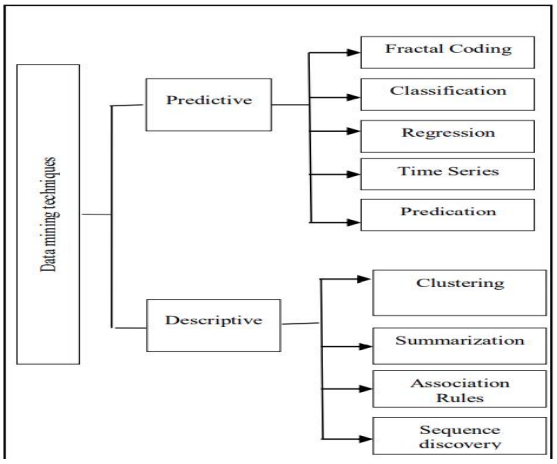


Fig. 1. Data mining techniques [34]

Because of the following, DM methods are successful and predictive of future patterns: a) it's easy to use, and it predicts outcomes based on previous events b) it operates by learning from previous data c) data from a variety of sources are managed, and only the information needed is extracted d) Relearning, past data, and evolving patterns are all easy ways to keep models up to date. This is what makes it dependable and realistic in the classification of medical images [36], [37], [38], [39].

4. DATA MINING FUTURE APPLICATIONS

In this part, we'll look at some of the DM applications and techniques.

1) DM applications in healthcare

Health DM tools have a lot of potentials and can be very useful. However, the availability of clean healthcare data is critical to the success of healthcare DM. In this regard, the healthcare industry must investigate how data can be collected, processed, prepared, and mined more effectively. Standardization of clinical terminology and data exchange across organizations are two possible directions for enhancing the benefits of healthcare DM applications [40].

1.1 Future directions of health care system through DM tools

Since healthcare data is not limited to quantitative data (e.g., doctor's notes or hospital records), It's critical to look at using text mining to expand the scope and scope of what healthcare data mining can currently do, according to the International Journal of Computer Science, Engineering and Information Technology (IJCSEIT). This is used to combine all of the data before mining the text. It's also worth investigating how images (such as MRI scans) can be incorporated into healthcare data mining applications. Progress has been made in these fields, it should be noted [40].

2) DM is used for the construction industry

The discovery of valuable knowledge from vast collections of data industries has drawn a lot of interest in the field of DM [41]. In the construction industry, DM from large amounts of data has become an essential method for information discovery. Energy, building occupant and occupancy actions, safety management, material efficiency, and textual information discovery are some of the most common DM application domains in the construction industry [42].

2.1 Future Directions of construction industry system through DM Tools

Two major developments in the building industry's future growth are sustainable construction and digital construction. Energy management, safety management, and green building all fall under the scope of sustainable construction in a broad sense. This shows that the use of data mining in sustainable construction is a hot topic, regardless of the past, current, or future [40].

3) DM methods are used in the web education

In the field of web education, DM techniques are used to upgrade courseware. The connections are discovered by looking at the consumption data collected during students' sessions. This expertise is extremely beneficial to the course's instructor or author, who can determine which changes are most necessary to increase the course's effectiveness. In the twenty-first century, beginners use DM techniques, which are one of the most powerful learning methods available. This allows learners to become more conscious of their surroundings. The application of DM techniques to educational chats is both feasible and can improve learning environments in the twenty-first century, according to Web Education [40].

4) In agriculture

Scientists and researchers around the world are dealing with how to make agriculture safe and resilient in the face of continuing conditions and environmental change. Transition and multidisciplinary approaches are needed in the agricultural system. For the production and efficiency when working with the same limited resources, intelligent and precision agricultural approaches were prioritized [43]. The strategy requires the collection of data from a variety of sources and the effective application of that data in the appropriate area. As a result of this need, there has been an increase in interest in extracting information from large troves of data resulting from various research and survey projects. DM techniques advanced the concept of knowledge generation and pattern recognition when they first appeared. Even though DM is a new science, it has a wide range of applications in agriculture and related industries, and it has a bright future [44].

5. RELATED WORK

Singh et. al. [45] compared the performance of different classifiers (DT classifier (J4.8, Simple CART)), (Bayes classifier (NB, Bayesian LR)). They were the most popular DM algorithms for BC classification. This paper aimed was to determine which classifier produces the most reliable results for the Wisconsin Breast Cancer (original) dataset WBCO. Dataset of BC was taken from the UCI ML repository using the WEKA tool. The experimental results show that the DT classifier, i.e., Simple CART (98.13 percent), has the highest accuracy of all the classifiers.

Bataineh et. al. [46] presented five nonlinear algorithms including K-NN, Multi-Layer Perceptron (MLP), Classification and Regression Tree (CART), Gaussian NB, and SVM was done for BC detection. The author's main goal was to compare the performance and efficacy of BC detection algorithms. The accuracy of each algorithm was also calculated separately by the author. A dataset of Wisconsin BC diagnostics was used in the study (WBCD). To calculate the accuracy of each algorithm, the author used the K Fold validation process. MLP outperformed the K-NN, CART, and NB algorithms with an accuracy of 96.70 percent.

Sinha et al. [47] introduced attribute filtering strategies, such as frequent itemsets mining, to identify the most important and applicable attribute from the Wisconsin BC dataset using a classification algorithmic such as SVM. Attribute filtering was used to compare NB, K-NN, and DT. With attribute filtering, SVM generated the highest area under the curve as compared to other classification techniques, resulting in better field accuracy and ROC curve.

Bharati et al. [48] presented the capability of the classification of NB, Random Forest (RF), LR, MLP, K-NN in evaluating the BC disease dataset from the UCI repository, which was observed to predict the existence of BC. The data set consisting of Kappa Statistics, TP Rate, FP Rate, and other metrics have all been investigated exactitude. The efficiency of the K-NN classifier algorithm was observed.

Ghani et al. [49] used anthropometric data and parameters obtained during routine blood processing that can be used to predict BC. Using the recursive feature elimination process, they first identified the most relevant attributes in the dataset that could be used as biomarkers. They discovered that the best biomarkers for BC are age, BMI, glucose, HOMA, and resistance. K-NN, ANN, DT, and NB classification techniques were used for classification. ANN was found to be the most accurate at classifying the attribute, with an accuracy of 80.00 percent.

Basunia et al. [50] proposed a stacking classifier ensemble approach that effectively classifies benign and malignant tumors by combining multiple classification techniques. Their experiment used the “Wisconsin Diagnosis BC”

dataset from the UC Irvine Machine Learning Repository. They chose 20 top features for BC prediction using the Univariate Feature Selection process. Jupyter Notebook is used with some Python open-source libraries to implement various classification techniques such as CART, LR, K-NN, SVM, RF, and Stacking Classifier techniques. The overall outcomes indicate Stacking classifier has the highest accuracy 97.20%.

Saoud et al. [51] used feature selection techniques to enhance the accuracy of six algorithms for BC classification and diagnosis: Bayes Network, SVM, K-NN algorithm, ANN, DT (C4.5), and LR. They used both databases WBC and WBCD. The feature selection technique increased the accuracy of some classifiers, such as BN, in both WBCD and WBC. However, some classifiers, such as SVM, had their accuracy reduced as a result of the feature selection technique. The BN with feature selection is the best model for classifying BC in WBC, while SVM without feature selection is the best for WBCD.

Kumar et al. [52] proposed two datasets of BC, taken from the UCI Machine Learning repository. On both datasets, seven algorithms were used. Which are (Bayes network, NB, SVM, K-NN, DT, RF, MLP). These two datasets have various features, with 11 and 32 features respectively. The datasets are split into two parts. The training data accounts for 65 percent of the overall dataset, while the evaluation data account for the remaining 35 percent. The accuracy of the Bayesian Network technique on the BCDW 11 dataset was 97.13 percent, while the SVM technique on the WBCD dataset was 97.89 percent.

Sakri et al. [53] proposed integrating the feature selection algorithm with classification algorithms in BC prognosis. They claimed that using feature selection techniques to reduce the number of features in most classification algorithms, can improve them. Some features are more significant and have a greater impact on the classification algorithms' results than others. They presented the results of their experiments with and without the feature selection algorithm, particle swarm optimization (PSO), on three common classifying algorithms, namely NB, K-NN, and REP Tree. As a result, NB obtained better results with and without PSO, while the other two techniques performed better with PSO.

Sudha et al. [54] suggested an improved lion optimisation algorithm (ILOA) technique that can identify small feature subsets quickly and accurately to classify the BC data set. A total of 500 mammogram images (288 benign and 212 malignant) were used as a case sample in this proposed study. After segmentation, each mass was represented with 123 features, including 96 texture features, 9 histogram features, 11 shape features, and 7 radial distance features, using a region growing algorithm. The Feature selection technique used a minimum distance classifier, K-NN classifier, and SVM classifier. As compared to other algorithms, ILOA with K-NN classifier performed well for BC classification [55].

6. MATERIALS AND METHODS

In this proposed model many classifiers were used to classify the Breast Cancer tumor with high accuracy, efficiency using via LR, fine K-NN, linear SVM, weighted K-NN, gaussian NB using nine features. The used dataset was the UCI breast cancer machine learning repository. The mechanism of this proposed model goes through five main stages, which are (Data Processing, Validation Choosing, Classification and Evaluating the results), as demonstrated in Fig. 2 that shows the Flowchart Diagram of the proposed model. The preprocessing method is done for missing feature values (in Single Epithelial Cell Size feature, there are 16 instances in Groups 1 to 6 that contain a single missing value in breast cancer dataset (i.e., (unavailable), attribute value denoted by “?”). To test the predictive accuracy of the fitted models, use the 10-fold cross validation process by MATLAB as a classifier tool in this study. After the classification results of all the five algorithms, the performance was measured by the confusion matrix and ROC area.

6.1 Dataset

The data for this study was provided by the UCI Machine Learning repository, which is located in the BC Wisconsin sub-directory, with 699 examples, two classes (malignant and benign), and 9 integer-valued attributes (as shown in Table 1). In UCI Breast cancer dataset (Dataset’s link:https://archive.ics.uci.edu/ml/datasets/breast +cancer) the class distribution are as following:

1- Benign class: 458 (65.5%) instances.
2- Malignant class: 241 are (34.5%) instances.

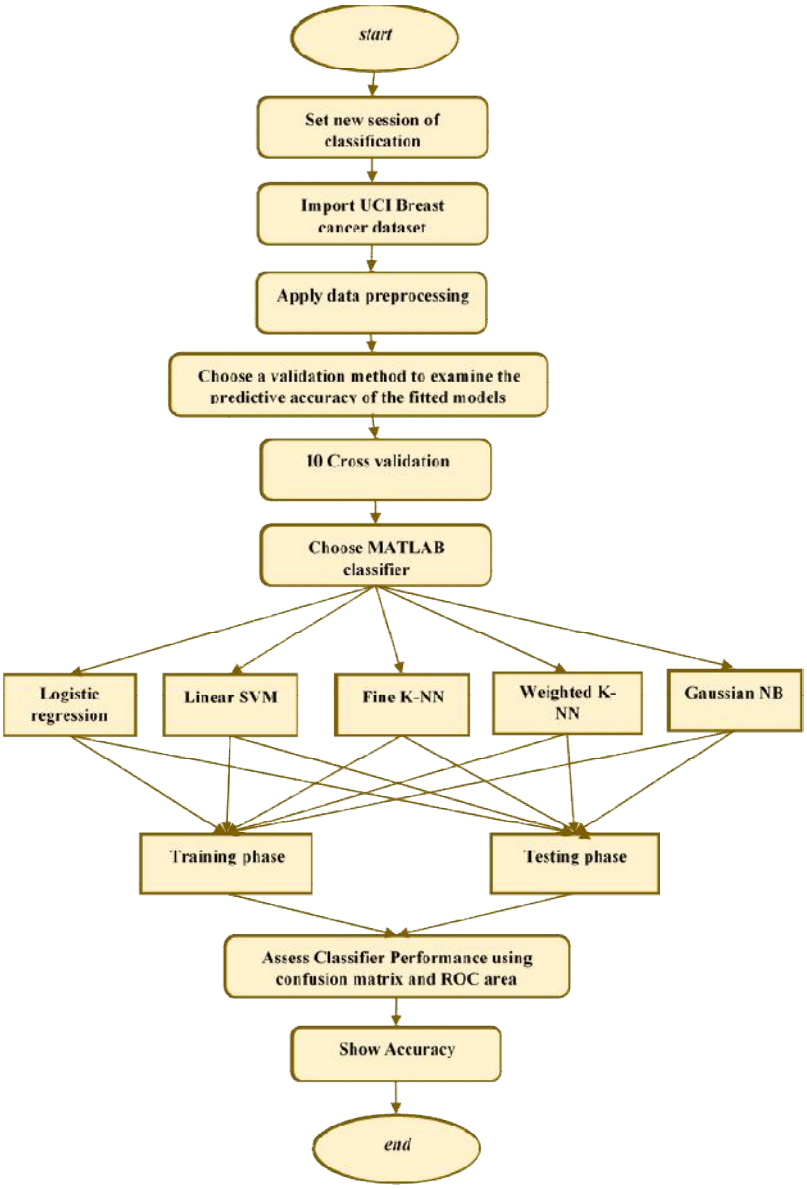


Fig. 2. The Proposed Model Flowchart Diagram

6.2 Classification Algorithmes

6.2.1 Support vector machine (SVM)

SVM is a superior DM technique that produces accurate classification results [56], [57], [58]. Only data sets with exactly two groups to classify can be used with SVM. It categorizes data by deciding the best hyperplane that separates all data points into one of two classes. SVM's main goal is to maximize the margins between two hyperPlane classes. Cancer diagnosis, Face recognition, and text categorization are examples of real-world applications of SVM. When dealing with binary classification, it is an effective technique. Can have the right to select the normalization of w>x + b = 0 and c (w>x + b) = 0 since they define the same plane. Select normalization such that positive and negative support vectors are w>x++b=+1 and w>x+b = 1, respectively [59], [60], [61].

The margin is then calculated as follows:

$$\frac{w}{\|w\|}(X_+ - X_-) = \frac{W^T(X_+ - X_-)}{\|W\|} = \frac{2}{\|W\|} \quad (1)$$

Table 1. Breast cancer dataset attribute information

Attribute	Domain
1) Sample code number	ID Number
2) Clump Thickness	1-10
3) Uniformity of cell size	1-10
4) Uniformity of cell shape	1-10
5) Marginal Adhesion	1-10
6) Single Epithelial cell size	1-10
7) Bare Nuclei	1-10
8) Bland Chromatin	1-10
9) Normal Nucleoli	1-10
10) Mitoses	1-10
11) Class:	2 For benign
	4 For malignant

6.2.2 Logistics Regression (LR)

One of the most widely utilized generalized linear models in DM is LR [62]. The probability of an outcome that can take two values from a collection of predictor variables is predicted using LR. LR is primarily used for predicting and calculating performance probabilities [63].

6.2.3 K-Nearest Neighbors (K-NN)

K-NN is a simple algorithm for instance-based learning that classifies objects in the feature space depending on their closest training dataset [64], [65]. An object is assigned to a class that includes its K-NN. A class is created for an object that includes its K-NN. To find the closest neighbor, the K-NN algorithm was used, which used Euclidean distance metrics [66], [67]. The equation below is used to measure the Euclidean distance metrics d(x,y) between two points x and y. w. Where N denotes the number of features with x = x1, x2, x3,...,xn and y = y1, y2, y3,...,yn [68].

$$d(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (2)$$

6.2.4 Weighted K-Nearest-Neighbors (Weighted K-NN)

This extension is based on the idea that observations in the learning set that are especially similar to the new observation (y, x) should be given more weight in the decision than observations that are far away from the new observation (y, x). This is not the case with K-NN: while only the k closest neighbors affect the prediction, this influence is consistent across these neighbors, despite their similarity to (y, x). To do this, the distances used in the first stage of the search for closest neighbors must be transformed into similarity measurements that can be used as weights [69]. Weighted K-NN assigns weights to each calculated value, then computes the nearest neighbors, and finally assigns the class to the processed instance [70], [71], [72].

6.2.5 Gaussian Naïve Bayes (Gaussian NB)

Gaussian NB algorithm uses for classification, which is a special form of NB algorithm [73]. When the features have continuous values or all of the features follow a Gaussian distribution such as a normal distribution, this method is particularly useful. The features' likelihood is assumed to be Gaussian [74].

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2{}_y}} exp(-\frac{(x_i-\mu_i)^2}{2\sigma^2{}_y} \quad (3)$$

In equation (3), x is a continuous data variable, and the parameters x and y are calculated using maximum likelihood estimation. After the data has been segmented by class, the mean μ_i and variance σ_y are measured.

6.3 The Evaluation Metrics of the Classifiers Performance

6.3.1 Confusion matrix

The confusion matrix (also called as the "Contingency Matrix") provides a good overview of the classifiers performance. Table 2 shows a standard confusion matric.

Table 2. A Typical 2x2 Confusion Matrix

		Predicted Class	
		Positive	**Negative**
Actual Class	Positive	TP	FN
	Negative	FP	TN

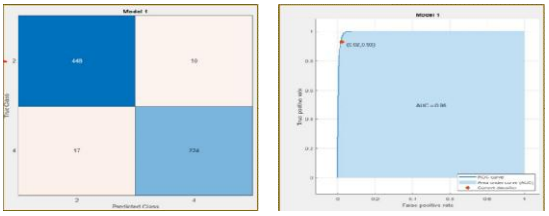
a. True positive (TP) : number of positive samples correctly predicted.
b. False negative (FN) : number of positive samples wrongly predicted.
c. False positive (FP) : number of negative samples wrongly predicted as positive.
d. True negative (TN) : number of negative samples correctly predicted.

6.3.2 Receiver operating characteristics area or ROC area

A ROC curve is a graphical representation of the true positive rate against the false-positive rate for different diagnostic test thresholds. The ROC curve is used to measure a classifier's performance and to give a higher score than the previous classifier. The false positive rate is known as specificity and the true positive rate is also known as sensitivity. Excellent (0.90-1), good (0.80-0.90), fair (0.70-0.80), bad (0.60-0.70), and fail (0.50–0.60) are used to evaluate a classifier's performance.

7. RESULTS AND DISCUSSION

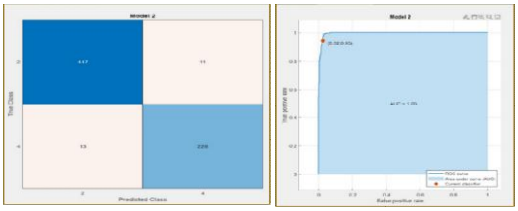
The research used the confusion matrix and area under the ROC curve, to determine the degree of performance and applicability of the models.



a. Confusion matrix b. ROC curve

Fig. 3. Evaluation results for logistic regression (Model 1)

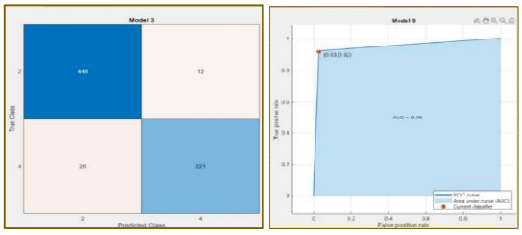
From Fig. 3 (a) the confusion matrix calculated for model 1(LR), shows the true and false positive and true and false negatives of the train set. In this confusion matrix, there are 448 patients, who have true benign lesions whereas 10 have false benign. Also have, 224 patients who have a truly malignant disease and 17 cases are false malignant. Fig. 3 (b) illustrates the ROC curve plot. The area under the curve (AUC) is 0.99 and is close to 1 (which is Excellent).



a. **Confusion matrix b. ROC curve**

Fig. 4. Evaluation results for SVM (Model 2)

Fig. 4 (a) shows the confusion matrix for SVM, there are 447 patients, which are true benign whereas 11 are false benign. Also had 228 true malignant and 13 false malignant patients. It also has a very good area under the ROC curve (AUC=1.00 the ideal value).



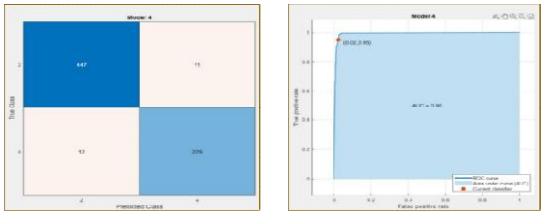
a. Confusion matrix b. ROC curve

Fig. 5. Evaluation results for K-NN (Model 3)

Moreover, the classification made by the K-NN was also evaluated using a confusion matrix Fig. 5 (a). In this matrix, there are 446 patients, which are true benign whereas 12 are false benign. Also have 221 patients who are true malignant and 20 false malignant patients. Fig. 5 (b) the ROC curve plot.The area under the curve (AUC) is 0.95 and is close to 1 (which is Excellent).

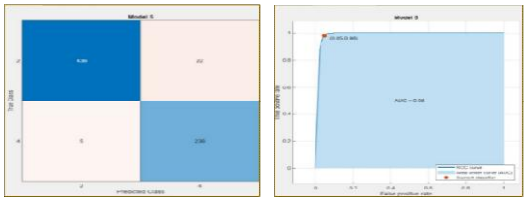
The confusion matrix for weighted K-NN is illustrated in Fig. 6 (a), there are 447 patients, which are true benign whereas 11 are false

benign. We also have 229 true malignant and 12 false malignant patients. Fig. 3 (b) illustrates the ROC curve plot. The area under the curve (AUC) is 0.99.



a. Confusion matrix **b. ROC curve**

Fig. 6. Evaluation results for Weighted K-NN (Model 4)



a. Confusion matrix **b. ROC curve**

Fig. 7. Evaluation results for Gaussian NB (Model 5)

Above in Fig. 7 (a) is the confusion matrix. There are 436 true benign patients in this matrix, while 22 are false benign. We also have 236 true malignant and 5 false malignant patients. Fig. 7 (b) is the ROC curve plot. The AUC measures the training accuracy. The (AUC) is 0.98 and is close to 1 (which is Excellent).

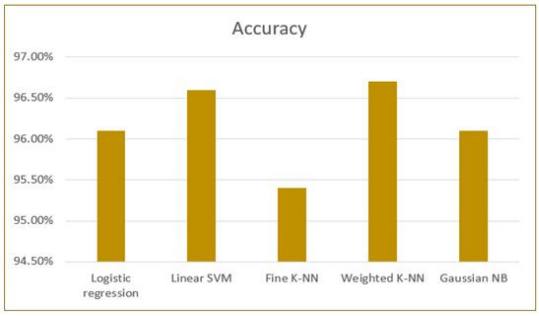


Fig. 8. Comparison of accuracies between all the classifiers

The overall outcomes displayed in Table 3 indicate that Weighted K-NN Classifier has the highest accuracy 96.7%, where Fine K-NN has the lowest accuracy 95.4%. Also, the results show can the Weighted K-NN Classifier has the best training time value (0.5096 sec) and ROC area 0.99. The Weighted K-NN method is the best classifier among the five proposed classifiers for classifying a tumor as benign or malignant, according to these findings.

As compared to other classifiers, weighted K-NN has the highest accuracy of 96.7 percent, as shown in Fig. 8.

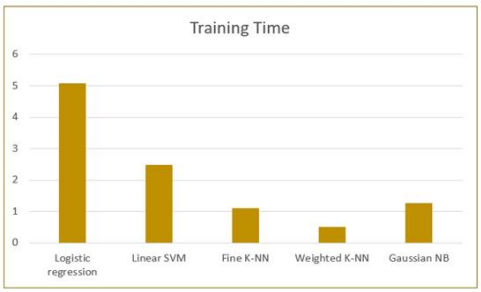


Fig. 9. Comparison of training time between all the classifiers

Fig. 9, Shows the training times of all the five classification algorithms. The training time for weighted K-NN is less than other algorithms.

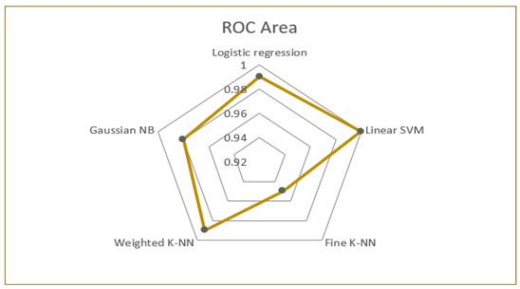


Fig. 10. ROC Area of all classifiers

From above figure show the graphical representation of the ROC area in MATLAB of the five classifiers on the dataset. ROC area of the linear SVM gave a better result, followed by the weighted K-NN, LR, Gaussian NB then fine K-NN classifiers.

8. Comparitive Study

The comparison summary of the related works is shown in Table 4. The researchers in the related papers used various techniques of feature selection and classification methods, as well as different datasets with different numbers of

Table 3. Performance Study of Algorithms

Model no.	Model Type	Accuracy	Training time (sec)	Area under ROC curve
1	Logistic regression	96.1%	5.0739	0.99
2	Linear SVM	96.6%	2.4899	1.00
3	Fine K-NN	95.4%	1.1064	0.95
4	Weighted K-NN	96.7%	0.5096	0.99
5	Gaussian NB	96.1%	1.2643	0.98

Table 4. Comparison of related works

R#	Classifier	Tool	Dataset	Number of attribute	Data type	Data processing method	Evaluation method	Validation technique	Accuracy
[45]	Naive Bayes Bayesian Logistic Regression Simple CART J48	WEKA	UCI repository	11 Attributes	Numeric (Discrete value)	Kappa statitics Mean Absolute error	Performane classifiers	-	95.26% 65.42% 98.13% 97.27%
[46]	MultiLayer preceptron K-Nearest Neighbours CART Gussian Naive Bayes Support vector machine	MATLAB	UCI repository WDBC dataset	32 Attributes	Images	Standardize rescaling method	Binary classification Accuracy method	cross validation	99.12% 95.61% 93.85% 94.73% 98.24%
[47]	Support vector machine Naïve Bayes k-Nearest Neighbours Decision Tree	PYTHON	UCI repository WBC	31 Attributes	Numeric (binary value)	z-score normalization	Confusion Matrix	-	96.61% 96.46% 91.74% 90.27%
[48]	K-Nearest Neighbors Naïve Bayes Random Forest Logistic Regression Multilayer Perceptron	WEKA	UCI repository	10 Attributes	Numeric	Kappa Statistics	Binary classification Accuracy method	-	72.37 % 71.67% 69.58 % 68.8% 64.68 %
[49]	K-Nearest Neighbors Artificial Neural Networks Decision Tree Naive Bayesian	WEKA	UCI repository	9 Attributes	Numeric	Recursive feature Elimination for features selection	confusion matrix	-	77.14% 80.00% 71.43% 73.91%
[50]	CART	PYTHON	UCI	32	Numeric	Features selection	confusion matrix	Cross	94.74%

R#	Classifier	Tool	Dataset	Number of attribute	Data type	Data processing method	Evaluation method	Validation technique	Accuracy
	Logistic regression K-Nearest Neighbors Support Vector Machine Random Forest Stacking Classifier		repository	Attributes				validation	97.08% 95.91% 95.91% 97.08% 97.20%
[51]	Bayes Network Support Vector Machine K-Nearest Neighbors Artificial Neural Networks Decsion Tree (C4.5) Logistic Regression	WEKA	UCI repository WBC WBCD	9WBC 32WBCD	Numeric	Features selection	confusion matrix	Cross validation	With WBC (BN):97.42% With WBCD (SVM): 97.36%
[52]	Bayesian network Naïve Bayes SVM Multi Layer preceptron K-NN Decision Tree (J48) Random Forest	WEKA	UCI repository BCWD WBCD	11BCWD 32WBCD	Numeric	Data statistics	Performance classifiers	Cross validation	With BCDW (Bayesian Network): 97.13% With WBCD (SVM): 97.89%
[53]	Naïve Bayes K-Nearest Neighbors Fast decision tree learner (REPTREE)	WEKA	UCI repository	35 Attributes	Numeric	Features selection and extraction	confusion matrix	Cross validation	81.3% 75.0% 93.6%
[54]	Support Vector Machine K-Nearest Neighbours	MATLAB	Digital database for screening mammography (DDSM)	-30 Attributes	Images	Features selection and extraction	Performance classifiers	Cross validation	98.92% 99.31%
Proposed Work	Logistic regression Support Vector Machine Weighted K-NN K-Nearest Neighbours Gaussian Naïve Bayes	MATLAB	UCI repository	9 Attributes	Numeric	-	Confusion matrix ROC area	Cross validation	96.1% 96.6% 96.7% 95.4% 96.1%

features. Comparing with the previous works, the provided method acquires a high accuracy classification of breast cancer. However, researchers in [47] and [45] used WBC (original) dataset to train and test different DM algorithms. They respectively registered anaccuracy of 96.61% (SVM) and 98.13% (CART), despite a high execution time of CART. Researchers in [46] used the WDBC dataset with the standardization method to reach 99.12% for MLP. In [48], researchers used fewer attributes and gained an average of 72.37% accuracy for K-NN, 71.67% for NB, 69.58% for RF, and 64.68% for LR. Researchers in [49] obtained 80% for ANN by using the feature selection method. In [50], researchers used the feature selection method to reach 97.20% for Stacking Classifier, researchers in [51] used two datasets with a feature selection technique to reach 97.42% from BN for WBC, 97.36% from SVM for WBCD. Researchers in [52] used two different datasets to reach a high accuracy rate of 97.13% from BN for the BCDW dataset, 97.89% from SVM for WBCD. In [53], researchers used many features but achieved fewer accuracy rates (93.6%) for DT. Lastly, researchers in [54] gained a good accuracy result (98.92%) for K-NN. The proposed work utilized five DM classifiers (Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), weighted K-Nearest Neighbors (Weighted K-NN), and Gaussian Naïve Bayes (Gaussian NB) algorithms) and the best classifier was Weighted K-NN with 96.7% accuracy.

9. CONCLUSION

This paper attempted to improve the accuracy of breast cancer classification using data mining techniques. In this study the UCI breast cancer dataset used and five data mining algorithms were used for the classification (Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), weighted K-Nearest Neighbors (Weighted K-NN), and Gaussian Naïve Bayes (Gaussian NB) algorithms). All the experiments were done using MATLAB 2021a. The primary goal is to assess how well each algorithm performs in terms of classification test accuracy when it comes to classifying data. The evaluation of the results done in terms of the confusion matrix and ROC curve. Investigational results show that the Weighted K-NN classifier has the highest accuracy 96.7%, where Fine K-NN has the lowest accuracy 95.4%. The last four classifiers respectively are Linear SVM, LR, Gaussian NB, and Fine K-NN with the accuracy ratios of 96.6%, 96.1%, 96.1%, and 95.4%.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Abdulqader DM, Abdulazeez AM, Zeebaree DQJML. Machine learning supervised algorithms of gene selection: A Review. 2020;62(03).
2. Ahmed O, Brifcani A. Gene expression classification based on deep learning. in 2019 4th Scientific International Conference Najaf (SICN). 2019;145-149:IEEE.
3. Zeebaree DQ, Haron H, Abdulazeez AM. Gene selection and classification of microarray data using convolutional neural network. in 2018 International Conference on Advanced Science and Engineering (ICOASE). 2018;145-150:IEEE.
4. Eesa AS, Abdulazeez AM, Orman ZJSJoUoZ. A DIDS based on the combination of cuttlefish algorithm and decision tree. 2017;5(4):313-318.
5. Taher KI, Abdulazeez AM, Zebari DAJAJoRiCS. Data mining classification algorithms for analyzing soil data. 2021;17-28.
6. Oskouei RJ, Kor NM, Maleki SAJAjocr. Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. 2017;7(3):610.
7. Ibrahim I, Abdulazeez AJJoAS, Trends T. The Role of Machine Learning Algorithms for Diagnosing Diseases. 2021;2(01):10-19.
8. Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed JJJoAS, Trends T. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. 2020;1(2):56-70.
9. Charbuty B, Abdulazeez AJJoAS, Trends T. Classification Based on Decision Tree Algorithm for Machine Learning. 2021;2(01):20-28.
10. Sagar M, Vivekkumar G, Reddy M, Devendiran S, Amarnath M. Research on intelligent fault diagnosis of gears using EMD, spectral features and data mining techniques. in IOP Conference Series: Materials Science and Engineering, 2017;263(6) :062047: IOP Publishing.

11. PadmaPriya R, Vadivu PSJIJoE, e-ISSN MR. A review on data mining techniques for prediction of breast cancer recurrence. 2019;2250-0758.
12. Zebari DA, Zeebaree DQ, Abdulazeez AM, Haron H, Hamed HNAJIA. Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. 2020;8:203097-203116.
13. Denny J, Ali S, Sobha TJIAJER. Efficient segmentation method for roi detection in mammography images using morphological operations. 2020;3(6):1-8.
14. Zeebaree DQ, Haron H, Abdulazeez AM, Zebari DA. Trainable model based on new uniform LBP feature to identify the risk of the breast cancer. in *2019* International Conference on Advanced Science and Engineering (ICOASE). 2019;106-111:IEEE.
15. Najat N, Abdulazeez AM. Gene clustering with partition around mediods algorithm based on weighted and normalized Mahalanobis distance. in 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). 2017;140-145: IEEE.
16. Khorshid SF, Abdulazeez AMJPsJoAoEE. Breast cancer diagnosis based on k-nearest neighbors: A review. 2021;18(4):1927-1951.
17. Zeebaree DQ, Haron H, Abdulazeez AM, Zebari DA. Machine learning and region growing for breast cancer segmentation. in 2019 International Conference on Advanced Science and Engineering (ICOASE). 2019;88-93:IEEE.
18. Zeebaree DQ, Abdulazeez AM, Zebari DA, Haron H, Hamed HNA. Multi-level fusion in ultrasound for cancer detection based on uniform lbp features.
19. Gupta S, Kumar D, Sharma AJIJoCS, Engineering. Data mining classification techniques applied for breast cancer diagnosis and prognosis. 2011;2(2):188-195.
20. Kadambari S, Jaswal K, Kumar P, Rawat S. Using twitter for tapping public minds, predict trends and generate value. in 2015 Fifth International Conference on Advanced Computing & Communication Technologies. 2015;586-589:IEEE.
21. El-Sebakhy EA, Faisal KA, Helmy T, Azzedin F, Al-Suhaim A. Evaluation of breast cancer tumor classification with unconstrained functional networks classifier. in IEEE International Conference on Computer Systems and Applications. 2006;281-287: IEEE.
22. Eesa AS, Brifcani AMA, Orman ZJIJoC, Engineering I. A New DIDS Design Based on a Combination Feature Selection Approach. 2015;9(8):1914-1918.
23. Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo EJAiim. A combined neural network and decision trees model for prognosis of breast cancer relapse. 2003;27(1):45-63.
24. Shrivastavat SS, Sant A, Aharwal RPJIJoACR. An overview on data mining approach on breast cancer data. 2013;3(4):256.
25. Moura DC, López MAGJIjocar, surgery. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. 2013;8(4):561-574.
26. Delen D, Walker G, Kadam AJAiim. Predicting breast cancer survivability: a comparison of three data mining methods. 2005;34(2):113-127.
27. O'Malley CD, Le GM, Glaser SL, Shema SJ, West DWJCIIJotACS. Socioeconomic status and breast carcinoma survival in four racial/ethnic groups: a population-based study. 2003;97(5):1303-1311.
28. Richards G, Rayward-Smith VJ, Sönksen P, Carey S, Weng CJAiim. Data mining for indicators of early mortality in a database of clinical records. 2001;22(3):215-231.
29. Shearer CJJodw. The CRISP-DM: the new blueprint for data mining. 2000;5(4).
30. Ramana BV, Babu MSP, Venkateswarlu NJIJoDMS. A critical study of selected classification algorithms for liver disease diagnosis. 2011;3(2):101-114.
31. Eesa AS, Orman Z, Brifcani AMAJEswa. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. 2015;42(5):2670-2679.
32. Sulaiman MAJJoSC, Mining D. Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. 2020;1(2):11-25.
33. Kumar AJSM, Structures. Different interface delamination effects on laminated composite plate structure under free vibration analysis based on classical

laminated plate theory. 2020;29(11): 115028.
34. Lashari SA, Ibrahim R, Senan N, Taujuddin N. Application of data mining techniques for medical data classification: a review. in MATEC Web of Conferences. 2018;150:06003:EDP Sciences.
35. Omar N, Abdulazeez AM, Sengur A, Al-Ali SGSJIJoEE, Science C. Fused faster RCNNs for efficient detection of the license plates. 2020;19(2):974-982.
36. Fayyad U, Piatetsky-Shapiro G, Smyth PJAm. From data mining to knowledge discovery in databases. 1996;17(3):37-37.
37. Hasan DA, Abdulazeez AMJI. A modified convolutional neural networks model for medical image segmentation. 2020;20:22.
38. Kareem FQ, Abdulazeez AM. Ultrasound medical images classification based on deep learning algorithms: A review.
39. Hiremath N, Reddy DMJMTP. Experimental studies to assess surface wear using grease degradation, bearing temperature and statistical parameter of vibration signals in a roller bearing. 2017;4(8):8370-8377.
40. Reddy DLCJIJoCA. A review on data mining from past to the future. 2011;975:8887.
41. Devendiran S, Mathew ATJMTP. Bearing Fault Diagnosis Using Empirical Mode Decomposition, Entropy Based Features And Data Mining Techniques. 2018;5(5):11460-11475.
42. Yan H, Yang N, Peng Y, Ren YJAiC. Data mining in the construction industry: Present status, opportunities, and future trends. 2020;119;103331.
43. Keleş MK. Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehnički Vjesnik. 2019;26(1):149-155.
44. Bhagawati K, Sen A, Shukla KK, Bhagawati RJIJoAER, Science. Application and Scope of Data Mining in Agriculture. 2016;3(7):236783.
45. Singh S, Thakral S. Using data mining tools for breast cancer prediction and analysis. in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018;1-4: IEEE.
46. Bataineh AAJIJoML,Computing. A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. 2019;9(3):248-254.
47. Sinha A, Sahoo B, Rautaray SS, Pandey M. Improved framework for breast cancer prediction using frequent itemsets mining for attributes filtering. in *2019* International Conference on Intelligent Computing and Control Systems (ICCS). 2019;979-982:IEEE.
48. Bharati S, Rahman MA, Podder P. Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. in 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). 2018;581-584: IEEE.
49. Ghani MU, Alam TM, Jaskani FH. Comparison of classification models for early prediction of breast cancer. in 2019 International Conference on Innovative Computing (ICIC). 2019;1-6: IEEE.
50. Basunia MR, Pervin IA, Al Mahmud M, Saha S, Arifuzzaman M. On Predicting and Analyzing Breast Cancer using Data Mining Approach. in 2020 IEEE Region 10 Symposium (TENSYMP). 2020;1257-1260:IEEE.
51. Saoud H, Ghadi A, Ghailani M, Abdelhakim BA. Using feature selection techniques to improve the accuracy of breast cancer classification. in The Proceedings of the Third International Conference on Smart City Applications. 2018;307-315:Springer.
52. Kumar A, Sushil R, Tiwari AJIJoCS, Engineering. Comparative study of classification techniques for breast cancer diagnosis. 2019;7(1):234-240.
53. Sakri SB, Rashid NBA, Zain ZMJIA. Particle swarm optimization feature selection for breast cancer recurrence prediction. 2018;6:29637-29647.
54. Sudha M, Selvarajan S, Suganthi MJIJoBIC. Feature selection using improved lion optimisation algorithm for breast cancer classification. 2019;14(4):237-246.
55. AK MF. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In healthcare. 2020;8(2): 111. Multidisciplinary digital publishing institute.
56. Han J, Kamber M, Pei JJTMKSiDMS. Data mining concepts and techniques third edition. 2011;5(4):83-124.
57. Cortes C, Vapnik VJMl. Support-vector networks. 1995;20(3):273-297.
58. Abdullah DM, Abdulazeez AMJQAJ. Machine Learning Applications based on

SVM Classification A Review. 2021;1(2):81-90.
59. Nisbet R, Elder J, Miner G. Handbook of statistical analysis and data mining applications. Academic Press; 2009.
60. Naveed N, Jaffar AJIJoPS. Malignancy and abnormality detection of mammograms using DWT features and ensembling of classifiers. 2011;6(8):2107-2116.
61. Kamalakannan J, Thirumal T, Vaidhyanathan A, MukeshBhai KD. Study on different classification technique for mammogram image. in 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], 2015;1-5:IEEE.
62. Tran HJn. Asurvey of machine learning and data mining techniques used in multimedia system. 2019;113:13-21.
63. Yusuff H, Mohamad N, Ngah U, Yahaya AJIJoR, Sciences RiA. Breast cancer analysis using logistic regression. 2012;10(1):14-22.
64. Purwanti E, Apsari R. Classification of digital mammograms using nearest neighbor techniques.
65. Mahmood MR, Abdulazeez AM. A Comparative study of a new hand recognition model based on line of features and other techniques. in International Conference of Reliable Information and Communication Technology. 2017;420-432:Springer.
66. Gareth J, Daniela W, Trevor H, Robert T. An introduction to statistical learning: With applications in R. Spinger; 2013.
67. Lavanya D, Rani DKUJIJoCS, Engineering. Analysis of feature selection with classification: Breast cancer datasets. 2011;2(5):756-763.
68. Medjahed SA, Saadi TA, Benyettou AJIJoCA. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. 2013;62(1).
69. Bhatia NJapa. Survey of nearest neighbor techniques; 2010.
70. Cherif WJPCS. Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. 2018;127:293-299.
71. Saeed J, Abdulazeez AMJJoSC, Mining D. Facial beauty prediction and analysis based on deep convolutional neural network: A review. 2021;2(1):1-12.
72. Bailey T, AK J. A note on distance-weighted k-nearest neighbor rules; 1978.
73. Witten IH, Frank EJASR. Data mining: practical machine learning tools and techniques with Java implementations. 2002;31(1):76-77.
74. Karabatak MJM. A new classifier for bresast cancer detection based on Naïve Bayesian. 2015;72:32-36.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/68450