Journal of Information Systems and Telecommunication
Vol.11, No.1, January-March 2023, 1-11
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Breast Cancer Classification Approaches - A Comparative Analysis

Mohan Kumar[1*], Sunil Kumar Khatri[2], Masoud Mohammadian[3]

[1]. Amity Institute of Information Technology, Amity University, Noida (UP), India
[2]. Director of Campus, Amity University Tashkent, Tashkent City, Uzbekistan
[3]. Faculty of Science and Technology, University of Canberra, Canberra ACT Australia

Received: 12 Jan 2022/ Revised: 30 Mar 2022/ Accepted: 26 Apr 2022

Abstract

Cancer of the breast is a difficult disease to treat since it weakens the patient's immune system. Particular interest has lately been shown in the identification of particular immune signals for a variety of malignancies in this regard. In recent years, several methods for predicting cancer based on proteomic datasets and peptides have been published. The cells turns into cancerous cells because of various reasons and get spread very quickly while detrimental to normal cells. In this regard, identifying specific immunity signs for a range of cancers has recently gained a lot of interest. Accurately categorizing and compartmentalizing the breast cancer subtype is a vital job. Computerized systems built on artificial intelligence can substantially save time and reduce inaccuracy. Several strategies for predicting cancer utilizing proteomic datasets and peptides have been reported in the literature in recent years.It is critical to classify and categorize breast cancer treatments correctly. It's possible to save time while simultaneously minimizing the likelihood of mistakes using machine learning and artificial intelligence approaches. Using the Wisconsin Breast Cancer Diagnostic dataset, this study evaluates the performance of various classification methods, including SVC, ETC, KNN, LR, and RF (random forest). Breast cancer can be detected and diagnosed using a variety of measurements of data (which are discussed in detail in the article) (WBCD). The goal is to determine how well each algorithm performs in terms of precision, recall, and accuracy. The variation of each classification threshold has been tested on various algorithms and SVM turned out to be very promising.

Keywords: Artificial Intelligence; Machine Learning; Wisconsin Breast Cancer Diagnostic (WBCD) Dataset; k-Nearest Neighbors (k-NN): Support Vector Classifier: Logistic Regression; ExtraTree-Decision; Random-Forest.

1- Introduction

One in eight women will be diagnosed with breast cancer throughout their lifetime, making it the leading cause of mortality in women[1]. Machine learning algorithms for early breast cancer diagnosis and prognosis have evolved as a result of improved technology and more access to data. Breast cancer is the largest cause of death for women, according to a WHO report, and the number of deaths is increasing every year [2]. Males are more prone than females to succumb to cardiac disease [3]. Breast cancer generally affects women over 40, but some risk factors make it possible for it to attack younger people as well [4]. It is impossible to make rational conclusions with traditional medical techniques and cannot handle enormous amounts of data without computational tools.

There are numerous systems that gather, analyse, and learn from cancer data, which explains why cancer data sources are so diverse. Even for data analysis, there are open-source tools that can be employed. The sheer volume of data involved makes data mining techniques very useful for early cancer detection. The earlier cancer is detected, the more accurate the diagnosis will be. For patients, improved therapy offers a higher probability of survival. Using previous data, classification models are built and trained for a given set of applications. This learning model reduces the amount of time and effort needed to manually predict and classify fresh data.

It also helps to detect difficult-to-perceive patterns from large and noisy datasets using machine learning algorithms by extracting knowledge and expertise from previous experiences. Experts believe that abnormal cell growth in the breast tissue is the cause of breast cancer in patients.

✉ **Mohan Kumar**
mohan_srivastava@hotmail.com

[5]. Figure 1 depicts a machine learning-based classification model for breast cancer. Before the characteristics can be extracted and the classification model constructed, the breast image database must be loaded. Benign non-cancerous conditions are those that are not life-threatening. Malignant cancer begins with abnormal cell proliferation and can spread quickly or even permeate neighboring tissues, making it somewhat lethal.[6].

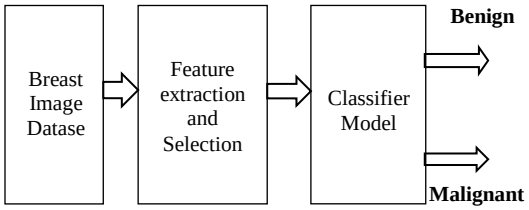


Fig. 1 Breast cancer image classification model [6]

It is hoped that the results of this study would help in the diagnosis of breast cancer by providing a classification model for malignant and benign breast cancers. Classification accuracy, Sensitivity, Specificity, and Precision are all assessed on the Wisconsin Breast Cancer Dataset (WBCD) [7] using k-nearest neighbors (k-NN), Logistic Regression, Support Vector Classifier (SVC), ExtraTree Decision (EDT), and Random Forest (RF). The most accurate models are LR, SVC, and DT, according to the results.. Seventy percent of the data is used for training, while the other thirty percent is used for testing. Data is divided in a 7:3 (70:30) ratio into training and testing portions because cross-validation is not a possibility here[8-9], the algorithms for classifying data performed admirably on the classification task.

The assessment of independent data set is analyses statistically using Cross validation technique for assessing. The evaluation of machine learning models is performed by model training on subsets of data input and available input data. Cross validation has a utility in scenario where there is a need of over-fitting. Optimal parameters are defined by cross validation. In current work cross validation not proved perfect for predicting over fitting in model. It was forcing model to get train again and again from scratch so current research work avoided the use of cross validation.

2- Literature Review

On this page, you'll learn about medically-approved machine learning algorithms used in breast cancer detection. Mangasarian and others [10] have used clinical data points to foresee the presence of cancer cells by forecasting the time interval at which recurrence will occur. Few of the studies conducted so far [11–13] also put forth work related to prediction of cancer and it's diagnosis by use of ML techniques such as decision tree to predict and diagnose cancer. According to Jin [14], KNN (K-Nearest Neighbor) is a classification algorithm used often among all machine learning approaches due to its simplicity and flexibility during implementation. The Wisconsin Prognostic Breast Cancer (WPBC) dataset was used by Belciug and colleagues [15] to compare K-means, cluster networks, and Self-Organizing Maps for the identification of breast cancer using the Wisconsin Prognostic Breast Cancer (WPBC) dataset [16], and the results showed that k-means were superior.

The Cancer Dataset was utilized by Chaurasia and Pal [17] prediction of breast cancer recurrence using ANNs, or artificial neural networks, Dyadic decision trees (DDTs), and Logistic Regression (LR) is implemented to retrieve accurate prediction about breast cancer. Research work outcome is the most accurate classifier for predicting the primary site of cancer development, Angeline [18] used the same dataset as Wisconsin Breast Cancer (WBC) and tested the performance of various algorithms such Nave Bayes, Decision tree (C4.5), KNN, and Support Vector Machine. When compared to the other approaches, Support Vector Machine performed the best. Author proposed and considered SVM as best classifier because it performs with high performance based on generalization as it doesn't add any prior knowledge. The function performed very well in high dimensional space of input dataset. This is the reason why SVM is being most promising as best classification function and has easily differentiate between instances of the two classes of breast cancer from training data.

Fuzzy clustering techniques were utilized by Abonyi and Szeifert [19] to predict cancer from the same dataset. the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). Lavanya and coworkers [20] employed a dynamic, hybrid technique to improve classification accuracy, and the dataset was cross-validated 10 times. The hybrid technique included various function in combination and results turned out very accurate and model performed well. Results from [21] research show that using a validation set to predict a model's performance in Machine Learning, cross-validation is the best technique. The cross-validation when applied to dataset proved very powerful tool. The data set is best utilized and shown comparably very good performance. There are some very complex machine learning models, cross validation help in using same data in each step of the pipeline.

For his studies, Cruz [22] investigated cancer detection and prognosis machine learning strategies. Using a microarray of cancer data, Tan A.C. [23] examined the bagged decision tree, also known as the C4.5 decision tree [24], he used bagging methods, decision trees, neural networks, and the Support Vector Machine to analyze a cancer database. According to the report's findings, bagging approaches have higher levels of accuracy. As evidenced

by the publications listed above, breast cancer can be diagnosed using a number of machine learning (ML) approaches. Comparing these approaches to discover the most efficient algorithm is difficult due to the use of various datasets and data sizes. The algorithms can only be employed on a single platform and foundation if they are implemented on a single dataset and with a single data size. This data will be used to develop the most accurate algorithm for classifying the presence of breast cancer. Our objective is to determine the most effective machine learning algorithm for the provided dataset.

3- Literature Review

Mammary cancer is classed as benign or malignant based on the results of tests such as the Komen Breast Cancer Early Detection and Management System (MBC-ELDS). Breast cancer is more accurately and efficiently classified as benign or malignant by employing the methods outlined above.

3-1- Proposed System Model

WBC: In this study, the dataset used was Wisconsin Breast Cancer, which can be obtained at the UCI Machine Learning Repository ((WBC). The pre-processed data will be fed into a variety of machine learning algorithms, including SVM, ExtraTree Decision, KNN, Random Forest, and Logistic Regression, to predict the likelihood of a cancer diagnosis. Measurement of performance and accuracy is the goal.

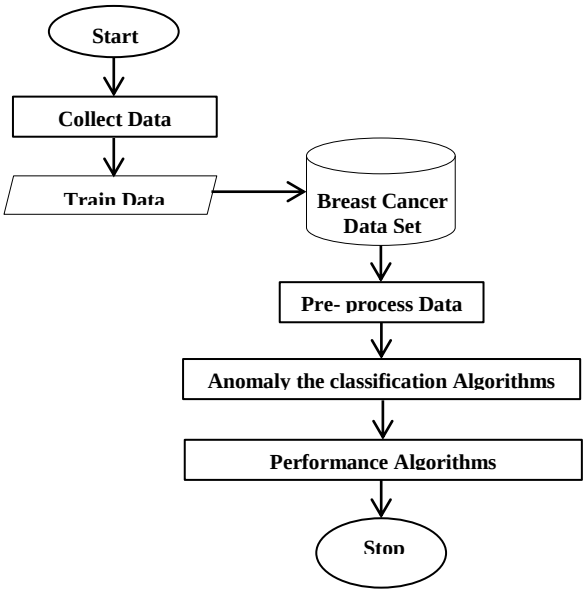


Fig. 2 Steps in proposed Methodology [25].

to do this, Machine Learning approaches make use of performance indicators such as accuracy, precision, sensitivity, and specificity. There will be a suitable categorization model proposed and validated. The proposed system model necessitates the completion of various steps. Fig.2 describes the method which has been conducted in the analysis.

3-2- Dataset

Diagnostic (WBCD) dataset, which was found in the UCI machine learning repository [7],[26]. For three years, researchers from the University of Wisconsin Hospitals (Wolberg, Street, and Mangasarian) collected this data.

We're utilizing the binary classification dataset for this exercise (WBCD). Within this collection, there are 569 records with 32 unique attributes. Breast lumps are aspirated using a fine needle aspirate (FNA) and then images of those aspirates are scanned for use in this dataset. The characteristics are as follows: Ten separate digitized cell nuclei features, mean and standard error, and the worst outcome for those ten diagnoses are included in these 30 real-value input feature attributes (Real, Positive). Type (B = benign, M = malignant) of cancer. Radius, Smoothness, Texture, Compactness, Fractal Dimension, Concavity, Perimeter, Convex Points, Symmetry, and Area can all be listed as qualities. [27]. The diagnoses were included in the dataset, which covered a wide spectrum of lumps and masses. A diagnosis is given when a tumor or lump is determined to be malignant or benign (B). There are four significant digits in these property values. Figure 3 reveals that 357 of the sample values are regarded as benign, whereas 212 are regarded as malignant.

Table 1: Wisconsin Dataset

Attributes	***Number of Attributes***
Sample Total	569
Dimensionality	30
Classes	2
Sample per class	Benign: 357 (62.74%) Malignant: 212 (37.26%)
Feature	Real Positive

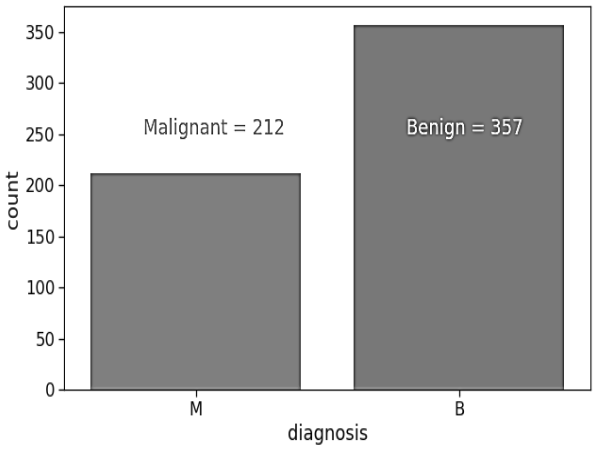


Fig. 3 Class distribution of WBCD data.

Heat maps in Figure 4 show the relationships between the variables. The correlation between two datasets shows the relationship between their attributes changing. A positive correlation occurs when two features move or change at the same time. If they go in the opposite direction, they are said to be negatively related. It reveals which traits are closely related to one another. It's critical to understand this since some machine learning algorithms do not work as expected.

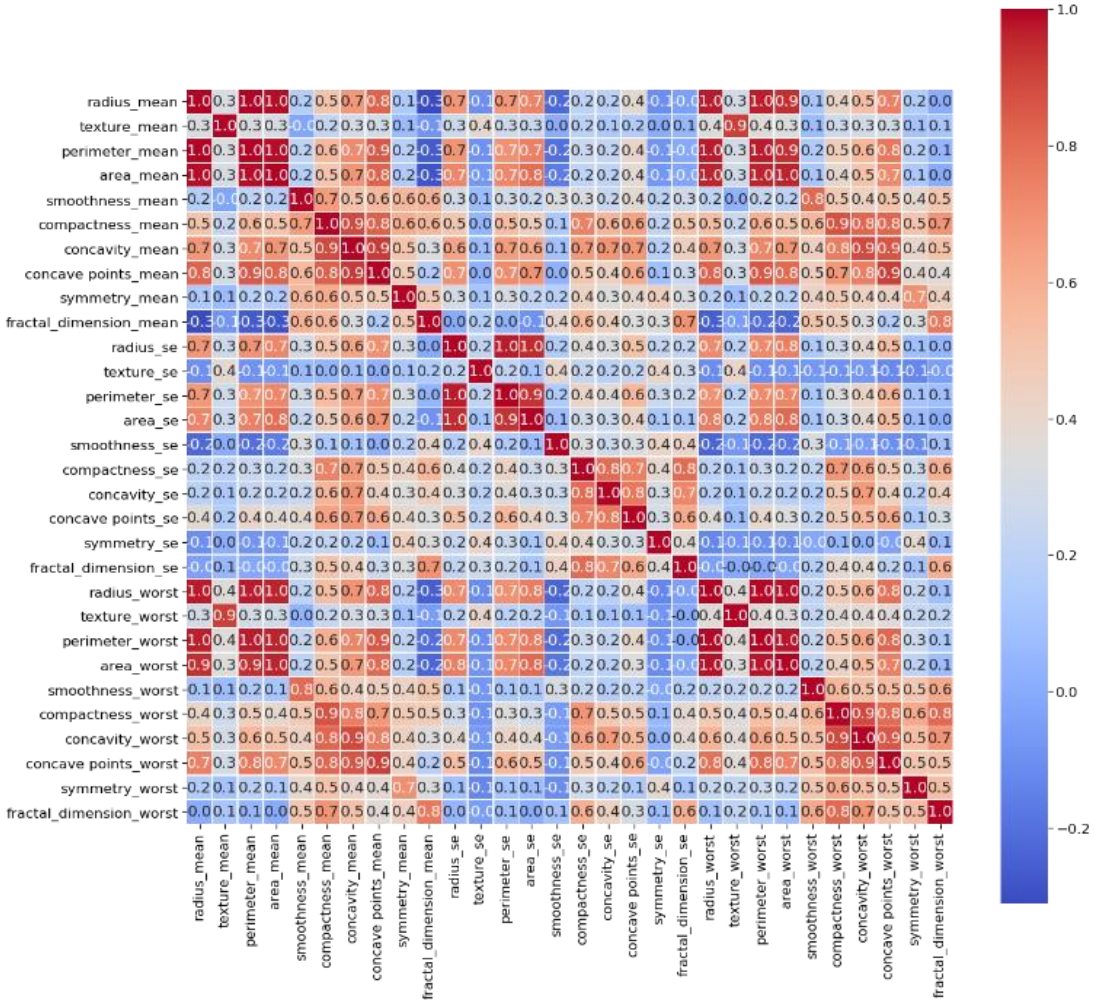


Fig. 4 Correlations between the features.

Many input features are highly linked in their data. The confusion matrix in Fig 4 showing collinear variables. Multi collinear variables are removed by few identified highly correlated variables which are independent. The combination of linear independent variables are summed up for performing analysis on variables which are highly correlated variables by applying principal components analysis or partial least squares regression.

These two ideas can be combined in a zillion different ways. Frequently, high-connectivity features are paired with ones that provide inaccurate information. By removing the most relevant features, we can get rid of the discriminating information they hold. We also see from this that while pushing birth control or other therapies we must remember that just because they can predict the outcome does not mean that they are at their core is an issue.

3-3- Preprocessing

Transformations are employed in pre-processing before the data is fed into the algorithm. the procedure used to collect the data Pre-processing is employed in order to turn unstructured data into something that can be read by machines. Standard methods for machine learning make the assumption that more data equates to better outcomes. In order to be used in this study, the WBCD data had to be normalized.

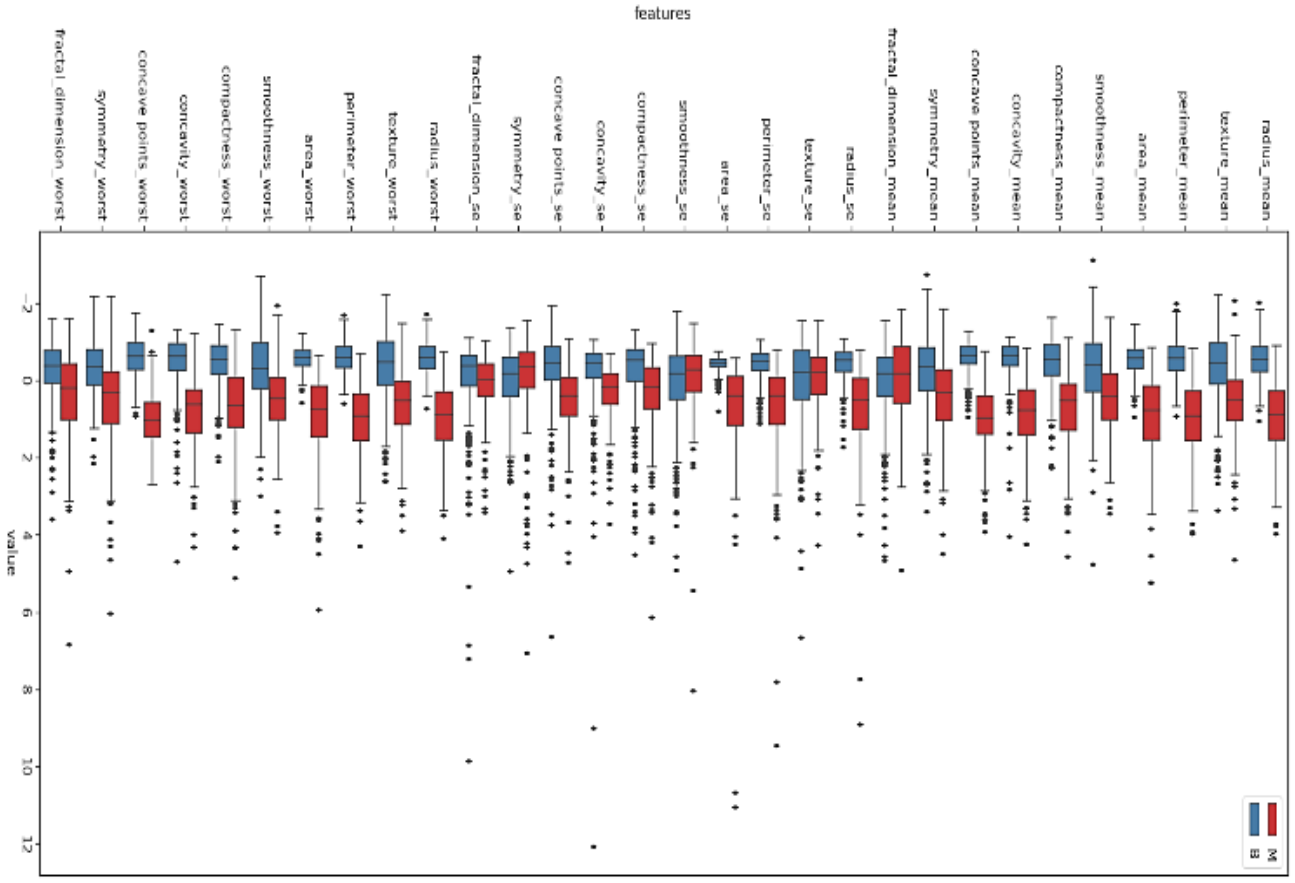


Fig. 5 Boxplot of features by diagnosis.

To produce the standard Gaussian distribution, several means with varying standard deviations are employed to turn the characteristics into a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. A distinct ID is assigned to each patient in the database to represent them in the dataset. The first attribute id, does not include any references to this variable. The next consideration is the diagnostic value. The recommended algorithms should be able to predict the value of this variable. A Z score denotes a feature's standardized value. How to compute Z-score: follow these steps.

$$Z = (x - \mu)/(\sigma) \quad (1)$$

x denotes the standardized value, whereas,, and the stddev are used to express the mean and standard deviation, respectively. After standardizing all numerical values, we create a box plot (Fig. 5) to organize the data. Because there were so many outliers in the data, numerous features had to be discarded. As a result, when analyzing and categorizing the boxplot, we failed to take advantage of these features. There are some data points in analysis which are very far from other data points.

The outlier retrieved is very much different from other observations on axis. There are measurements which are very much different. There are experimental errors in the form of outliers. The analysis become very difficult when we try to apply statistic only because of outliers. So sometimes it is much needed to avoid them.

This chart shows how similar the graphics are amongst several of the features. There appears to be some similarity when comparing the perimeter with the area or the perimeter with the area se. However, certain traits differ based on whether a tumor is diagnosed as malignant or benign. Examples of the same include the concavity mean, radius means, and area means. However, if we look at metrics like texture se or fractal dimension mean, we can detect similarities in the distribution of malignant and benign tumors.

4- Nonlinear Machine Learning Algorithms

There are two types of machine learning algorithms: supervised and unsupervised. When using supervised learning, we provide a dataset to the machine to train it, and then the program outputs the accurate answers. Unsupervised learning, on the other hand, does not have any predetermined data sets or goals [28]. Machine learning has gained remarkable practical success in a variety of different fields. Medical expert systems [29], [30], social networks [31], and speech recognition [32], [33] are a few of the most common applications. As a result of past observations, machine learning learns how to make exact predictions. Classification is a supervised learning-related topic of study. The goal, as the name suggests, is to classify the samples in order to come up with a collection of plausible classes that are discrete. These include k-nearest neighbors (k-NN), support vector classifier (SVC), logistic regression, and extra-tree decision) and random-forest (RF) [34].

Classification involves educating the classifier to be aware of the different classes. As an added bonus, categorization is basically a mission to anticipate the value of a specific discrete characteristic. When it comes to classification, supervised learning algorithms are the go-to solution. This data is collected for the computer to learn a classification model from. After that, future projections will be possible. The classification problem is best solved with an algorithm based on supervised learning. Five different classification algorithms were used to sort the Wisconsin dataset [35]. To sum up, here are the five methods

4-1- K-Nearest Neighbors (K-NN)

It's possible to make predictions with supervised learning, but not with unsupervised learning. Depending on your requirements, it can be used for regression or clustering. The number of nearest neighbors is denoted numerically by a letter like K. There doesn't appear to be any formal training programme in place, as far as I'm aware. As part of the forecasting process, the Euclidean distance between the k-nearest neighbors is first taken into account [36].. When it comes to breast cancer, we've already labelled the cases as either malignant or benign. Labels are classified according to how far apart they are from one another. Figure 6 explains the KNN algorithm's technique better than any other figure.

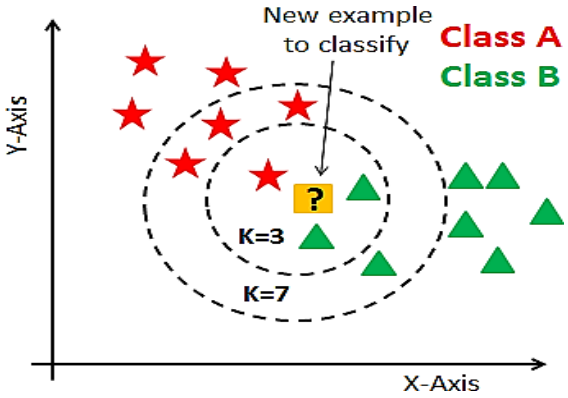


Fig. 6 Representation of KNN algorithm [26].

4-2- Logistic Regression (LR)

Logistic regression was first used in biology study in the early twentieth century. Using it in social studies classes is

becoming more and more common. This is also covered by predictive analysis. When there is only one dependent variable, it is most often used. To distinguish between a linear and a logistic regression, we employ the dependent variable. For continuous variables, the most common approach is linear regression, as shown in Figure 7. The following section outlines the many steps involved:

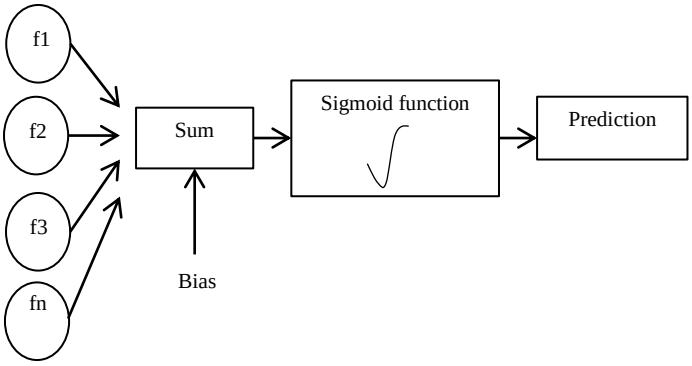


Fig. 7 Visualization of logistic regression steps.

Logistic regression makes use of both forward and backward propagation. Weight and characteristics are multiplied to begin the forward propagation phase. There's no way to tell how much something weighs when it's in its infancy. Because of this, weights can be assigned to certain random values. For sigmoid functions, the probability range is 0 to 1. According to the set threshold value, the forecasting method is carried out. A comparison of the projected and actual values is performed following the completion of the forecasting process. It's possible to calculate the difference between the projected and actual values using a loss function. The backward propagation approach is utilized if the loss function value is extremely high. By using backward propagation [38], one can obtain a derivative and update the weight values in accordance with the cost function. The sigmoid function is depicted in the following diagram:

$$\text{sigmoid}(x) = \frac{e^x}{1 + e^x} \quad (2)$$

4-3- Extra Tree-Decision (ETD)

Classifiers like the Extra Tree classifier are useful for combining the results of different decision trees that aren't related in any way. Foresters collect and sort these twigs to create different kinds of trees in the area. All of the decision trees have benefited from the use of genuine training data. If the loss function value is too large, the backward propagation strategy is used. A derivative is obtained and weight values depending on the cost function are updated using backward propagation [38]. Because of this, every call tree must start with the most basic features, which are then segregated from the data using statistical criteria known as the Gini Index. This random selection of options leads to a number of decision trees with no correlation. [39].

4-4- Support Vector Classifier (SVC)

SVMs, or Support Vector Machines, are frequently employed in machine learning. An N-dimensional hyper plane was sought for by these algorithms in order to classify the incoming data. Many hours are devoted to devising a plan that will maximise profits. When employing feature numbers, it is possible to distribute the N dimensionally uniformly. You can easily compare the two features. Despite the abundance of tools available, categorising things isn't always straightforward. To get the most accurate forecast, we should try to widen the margin of error. [40] This may be seen in Figure 8 where it appears that the SVM is:

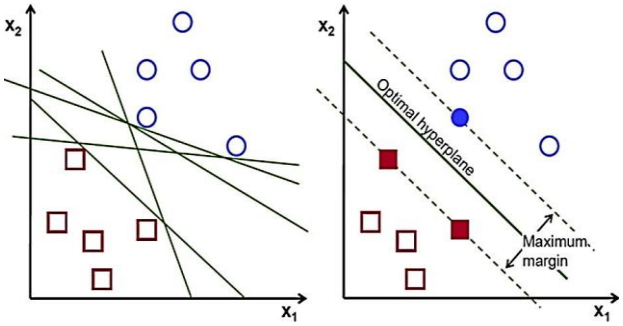


Fig.8 Support vector visualization [41].

Being over-classified often means being under-classified, and this must always be remembered. The error margin could be exceedingly tiny if we choose a categorization algorithm that does not sacrifice any of the individual samples. Final results may be less precise as a result of this situation. Support vectors that are most closely related to the hyper plane can also be considered as long as the gap between classes is as big as possible.

4-5- Random Forest (RF)

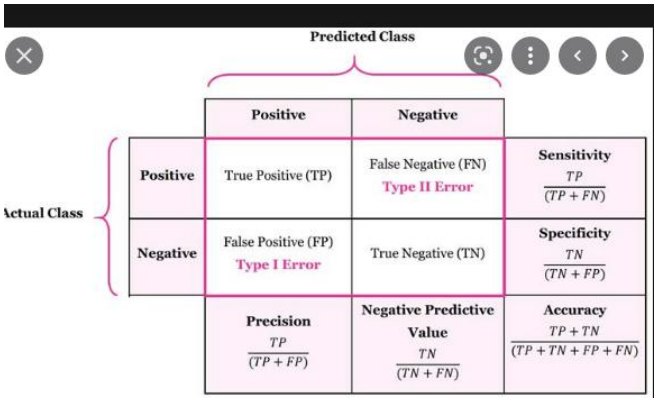
After the random forest, we come to an ensemble-learning model called the Random forest. This data analysis technique can be used for both regression and classification. It's termed a random forest since it's made up of many different decision trees. That is why utilizing a random forest instead of the more traditional decision tree seems sensible in some circumstances. It is necessary to create a classifier that uses the process of extracting these features[42], given that the rotating forest technique is being applied. After that, this attribute collection is divided up into K different groups using a randomization algorithm. Classifiers gain significance and precision in this way. [43].

The decision tree's most difficult element is selecting characteristics, and there are several ways to go about it.

Instead than seeking the most essential feature when dividing nodes, random forest seeks for the best features within a completely random collection of attributes. Even if it's curated even more haphazardly in the future, it's still doable. So it's plausible since, while using Random Forest, random attributes might be sought out rather than the most important traits for node splitting among the better ones. [44]. An internal function parameter can increase model accuracy or speed. The accuracy of the prediction model can be improved by utilizing characteristics such as Max features, minimum sample leaves, and n estimators, among others. When running the models, we frequently use N jobs and a randomized state machine in order to speed them up (RSM). Researchers employed characteristics including n estimators and random state parameters to calculate how many trees should be planted to improve the model's accuracy and speed.

5- Proposed System Model

Evaluation of the Model: Several performance measures can be used to assess the accuracy, sensitivity, specificity, and precision[45] of each Machine Learning approach. There are four ways to look at these choices: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [46]. The true positive decision happens when the instances of malignancy are predicted aptly. True negative decisions are an occurrence that occurs when instances of benign behavior are correctly predicted. In cases where a benign disease is predicted to be malignant, the False Positive decision is made [47]. A false negative occurs when a malignant disease is predicted to be benign [46].



i. Accuracy can be calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

ii. Sensitivity can be calculated as:

$$\text{Sensitivity (recall)} = \frac{TP}{TP+FN}$$

iii. Specificity can be calculated as:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

iv. Precision can be calculated as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

After these variables (Accuracy, Sensitivity, Specificity, and Precision) have been determined, a confusion matrix is created utilizing these values. When there is rise in classification threshold it will decrease the false positives, which further increased the precision. When there is decrease in raising the threshold of classification reduce the false positives which will further increase precision.

6- Result and Discussion

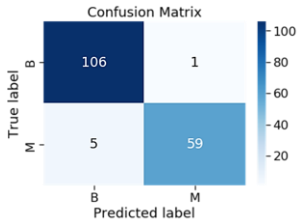
About 70% of the dataset items are used for training, while the remaining 30% are omitted for testing. Using a random number generator, these two subgroups are created from the entire dataset. Four separate measures are used to evaluate the performance of machine learning algorithms: specificity, sensitivity, precision, and accuracy (or accuracies). To determine the algorithm's overall performance, various measurements are compared. Multiple machine learning algorithms used to classify breast cancer data as benign or malignant, and then present the results in a confusion matrix (see Table II). Table III summarizes the various machine learning approaches to the breast cancer dataset based on various performance criteria. Table III. k-NN, support vector classifier, logistic regression, additional tree decision and random forest classification outcomes were evaluated. This highlights the areas of interest for various classification techniques. [48].

Tables III show that the SVC and ExtraTree classifiers are the most precise, specific, and accurate. On the other hand, radio frequency (RF) technology boasts the highest sensitivity. Results show that SVM and ExtraTree models identify tumors as benign or malignant with the greatest accuracy.

Table 2: Confusion Matrix.

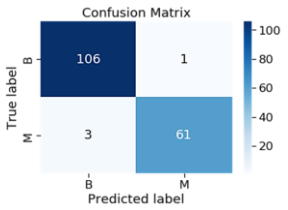
(i) KNN

KNN	Test Outcome	
Predictions	Benign	Malignant
Benign	106(TP)	1(FN)
Malignant	5 (FP)	59 (TN)



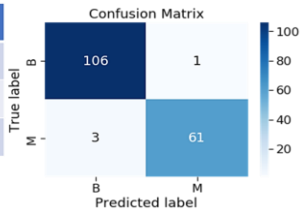
(ii) SVM

SVM	Test Outcome	
Predictions	Benign	Malignant
Benign	106(TP)	1(FN)
Malignant	3 (FP)	61 (TN)



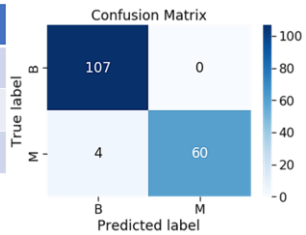
(iii) ExtraTree

ExtraTree	Test Outcome	
Predictions	Benign	Malignant
Benign	106(TP)	1(FN)
Malignant	3 (FP)	61 (TN)



(iv) Random Forest

Random Forest	Test Outcome	
Predictions	Benign	Malignant
Benign	107(TP)	0(FN)
Malignant	4 (FP)	60 (TN)



(v) Logistic Regression

Logistic Regression	Test Outcome	
Predictions	Benign	Malignant
Benign	102(TP)	5(FN)
Malignant	3 (FP)	61 (TN)

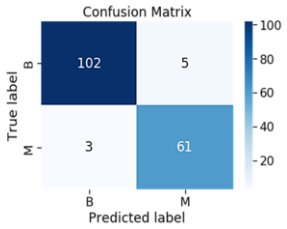


Table 3: Algorithm's comparison

	KNN		SVM	ExtraTree	RF	LR
Accuracy (%)	96.4		**97.7**	**97.7**	**97.7**	95.3
Sensitivity (%)	99.0		99.0	99.0	**100**	95.3
Specificity (%)	92.1		**95.3**	**95.3**	93.7	95.3
Precision (%)	95.4		**97.2**	**97.2**	96.3	97.1

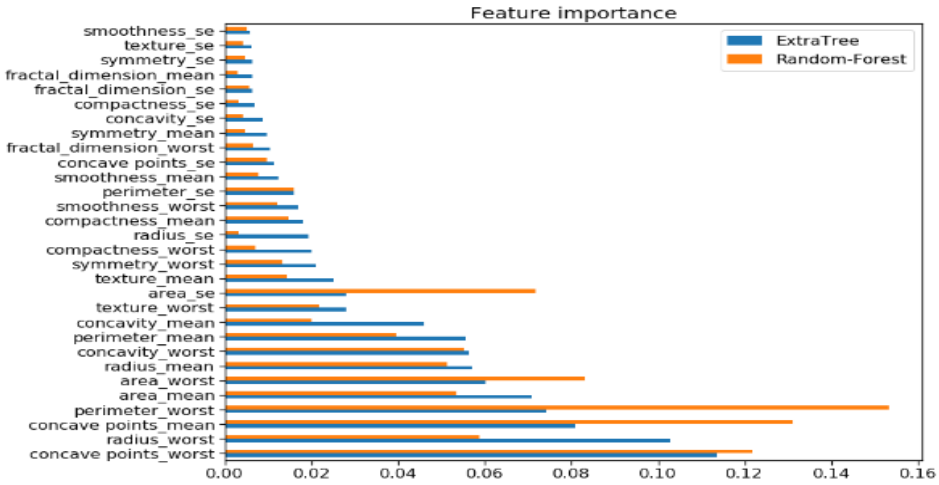


Fig.9 Feature Importance

Figure 9 represents feature selection which is most important for analyzing various machine learning algorithms

Table 4: Important Features

perimeter_worst	0.153370
concave points_mea	0.131221
concave points_worst	0.121852
area_worst	0.082954
area_se	0.071767
radius_worst	0.058753

The six most important features having data type float64 are mentioned on the above table

7- Conclusion and Future Scope

A number of machine learning algorithms, such as SVM, ExtraTree, Random Forest, and KNN, are put to the test. The analysis also makes use of Logistic Regression. Using a variety of machine learning methods, the goal was to discover the most accurate classifier. To sort the breast cancer dataset, all of the classifiers will be put to work (Wisconsin Breast Cancer Dataset). A Python tool was used to do the research. The accuracy, sensitivity, specificity, and precision of SVM and ExtraTree were superior to those of the other classifiers. Besides highlighting the six most essential qualities, the research also discusses how these characteristics affect the prediction models. We'll look into it some further. Alternative dataset combinations (such as 50-50, 80-20, 60-40, and so on) will be examined in a future study to see how well Bayesian Networks, Gradient Descents, and artificial neural networks perform when dimension reduction approaches are applied. The capabilities of KERAS-Neural-Networks Tensor Flow will be evaluated in the following publication by using deep learning that can handle enormous datasets, we'll see how deep learning compares to other approaches for handling large datasets. Because of this, knowing which method is best for which type of data will be much easier.

References

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," Cell, vol. 144, no. 5, pp. 646–674, Mar. 2011,

[2] S. Katuwal, P. Jousilahti, and E. Pukkala, "Causes of death among women with breast cancer: A follow- up study of 50 481 women with breast cancer in Finland," Int. J. Cancer, vol. 149, no. 4, pp. 839–845, Aug. 2021

[3] S. F. Khorshid and A. M. Abdulazeez, "breast cancer diagnosis based on k-nearest neighbors: A review," PalArch's J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 1927–1951, 2021

[4] Tawam Hospital |Medical News. (n.d.). Retrieved November 19, 2014, fromhttp://www.tawamhospital.ae/english/news/print.aspx?NewsID= 367

[5] M. Karabatak, "A new classifier for breast cancer detection based on Naive Bayesian," Measurement, vol. 72, pp. 32–36, 2015

[6] A-Al. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: a survey," Comput. Math. Methods Med., 2017

[7] M. Kumar, S. K. Khatri, and M. Mohammadian, "Breast cancer identification and prognosis with machine learning techniques-An elucidative review," J. Interdiscip. Math., vol. 23, no. 2, pp. 503–521, 2020,

[8] A. Joshi and A. Mehta, "Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer," Int. J. Emerg. Technol., vol. 8, no. 1, pp. 522–526, 2017.

[9] B. Soni, A. Bora, A. Ghosh, and A. Reddy, "RFSVM: A Novel Classification Technique for Breast Cancer Diagnosis," Int. J. Innov. Technol. Explor. Eng., 2019.

[10] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," Oper. Res., vol. 43, no. 4, pp. 570–577, 1995

[11] Zhi-H. Zhou and Y. Jiang, "Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble," IEEE Trans. Inf. Technol. Biomed., vol. 7, no. 1, pp. 37–42, 2003,

[12] T. Ornthammarath, "Artificial neural networks applied to the seismic design of deep tunnels," Università degli Studi di Pavia, 2007.

[13] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, no. 2, pp. 113–127, Jun. 2005

[14] Jini Marsilin, "An Efficient CBIR Approach for Diagnosing the Stages of Breast Cancer Using KNN Classifier," Bonfring Int. J. Adv. Image Process., vol. 2, no. 1, pp. 01–05, Mar. 2012

[15] S. Belciug, A-B Salem, F. Gorunescu, and M. Gorunescu, "Clustering-based approach for detecting breast cancer recurrence," in 2010 10th International Conference on Intelligent Systems Design and Applications, pp. 533–538, Nov. 2010,

[16] M.Lichman, "UC Irvine Machine Learning Repository," 2015. http://archive.ics.uci.edu/ml.

[17] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability," Int. J. Comput. Sci. Mob. Comput., vol. 3, no. 1, pp. 10–22, 2014.

[18] Christobel, Angeline, and Y. Sivaprakasam, "An empirical comparison of data mining classification methods," vol. 3, no. 2, 2011

[19] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," Pattern Recognit. Lett., vol. 24, no. 14, pp. 2195–2207, 2003

[20] D. Lavanya, "Ensemble Decision Tree Classifier For Breast Cancer Data," Int. J. Inf. Technol. Converg. Serv., vol. 2, no. 1, pp. 17–24, Feb. 2012

[21] Abad, Monica, James Carlisle Genavia, Jaybriel Lincon

Somcio, and Larry Vea. "An Innovative Approach on Driver's Drowsiness Detection through Facial Expressions using Decision Tree Algorithms." In 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0571-0576. IEEE, 2021.

[22] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Inform., vol. 2, p. 117693510600200030, 2006

[23] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," 2003

[24] G. L. Tsirogiannis, D. Frossyniotis, J. Stoitsis, S. Golemati, A. Stafylopatis, and K. S. Nikita, "Classification of medical data with a robust multi-level combination scheme," in 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2004

[25] G. Sudhamathy, M. Thilagu, and G. Padmavathi, "Comparative analysis of R package classifiers using breast cancer dataset," Int J Eng Technol, vol. 8, pp. 2127–2136, 2016.

[26] Frank A. UCI machine learning repository. http://archive. ics. uci. edu/ml. 2010.

[27] WHO, WHO position paper on mammography screening. World Health Organization, 2014.

[28] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in 2016 5th international conference on electronic devices, systems and applications (ICEDSA), pp. 1–4, 2016

[29] I. Castiglioni et al., "AI applications to medical images: From machine learning to deep learning," Phys. Medica, vol. 83, pp. 9–24, Mar. 2021

[30] R. Abdlaty, J. Hayward, T. Farrell, and Q. Fang, "Skin erythema and pigmentation: a review of optical assessment techniques," Photodiagnosis Photodyn. Ther., vol. 33, p. 102127, Mar. 2021,

[31] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," Heal. Inf. Sci. Syst., vol. 6, no. 1, p. 8, Dec. 2018

[32] M. D. Skowronski and J. G. Harris, "Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition," J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1817–1833, Mar. 2006

[33] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview," IEEE Trans. Audio. Speech. Lang. Processing, vol. 21, no. 5, pp. 1060–1089, May 2013

[34] "Cardio-Vascular Disease Prediction based on Ensemble Technique Enhanced using Extra Tree Classifier for Feature Selection," Int. J. Recent Technol. Eng., vol. 8, no. 3, pp. 3236–3242, Sep. 2019

[35] Kaggle.com, "Breast Cancer Wisconsin (Diagnostic) Data Set," 2021. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

[36] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 5, pp. 1774–1785, May 2018

[37] M. Faisal, A. Scally, R. Howes, K. Beatson, D. Richardson, and M. A. Mohammed, "A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation," Health Informatics J., vol. 26, no. 1, pp. 34–44, Mar. 2020

[38] V. Tatan, "Your Beginner Guide to Basic Classification Models: Logistic Regression and SVM," 2019.

[39] O. Maier, M. Wilms, J. von der Gablentz, U. M. Kramer, T. F. Munte, and H. Handels, "Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences," J. Neurosci. Methods, vol. 240, pp. 89–100, 2015

[40] S. B. Kotsiantis, "Decision trees: a recent overview," Artif. Intell. Rev., vol. 39, no. 4, pp. 261–283, 2013

[41] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA. Cancer J. Clin., vol. 68, no. 6, pp. 394–424, Nov. 2018

[42] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," Expert Syst. Appl., vol. 134, pp. 93–101, Nov. 2019

[43] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 10, pp. 1619–1630, 2006

[44] M. Maria and C. Yassine, "Machine learning based approaches for modeling the output power of photovoltaic array in real outdoor conditions," Electronics, vol. 9, no. 2, p. 315, 2020

[45] S. S. Shajahaan, S. Shanthi, and V. ManoChitra, "Application of data mining techniques to model breast cancer data," Int. J. Emerg. Technol. Adv. Eng., vol. 3, no. 11, pp. 362–369, 2013.

[46] S. Raschka, "An overview of general performance metrics of binary classifier systems," arXiv Prepr. arXiv1410.5330, 2014

[47] Madooei, Ali, Ramy Mohammed Abdlaty, Lilian Doerwald-Munoz, Joseph Hayward, Mark S. Drew, Qiyin Fang, and Josiane Zerubia. "Hyperspectral image processing for detection and grading of skin erythema." In Medical Imaging 2017: Image Processing, vol. 10133, pp. 577-583. SPIE, 2017.

[48] R. Abdlaty et al., "Hyperspectral Imaging and Classification for Grading Skin Erythema," Front. Phys., vol. 6, Aug. 2018