International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Breast Cancer Detection Using Machine Learning

Samarth Kumar[1], Simar Ahuja[2], Shobha Rekh[3]

[1,2]Student, Vit, Vellore
[3]Professor, Vit, Vellore

ABSTRACT
Breast cancer is one of the worst cancer diseases in the world which affects a very large number of people. Detecting and diagnosing this disease early is vital, since it is crucial for the survival of the patient. The main objective of this study is to investigate the different technologies that are used for identifying and diagnosing breast cancers. Furthermore, it explores different machine learning approaches, computer-aided detection systems (CAD) and common medical treatments used to diagnose and treat these fatal diseases on a higher level. Also, there is a difference between commercial software and tools that are used for identifying and assessing all stages of breast cancer against tools and software that are provided by nonprofit organizations. The conclusion of this report is that there is no method or standardized procedure to identify and diagnose breast cancer using all technologies**.**

Keywords: Breast Cancer, Dataset, CNN, KNN, Naïve Bayes, Random Forest, SVM, Logistic Regression.

1. INTRODUCTION

Next to lung cancer as the most fata disease, breast cancer is the most dangerous disease for a women all over the world. Based on the reports from WHO, this disease becomes the first widely occurring cancer, among women, worldwide. It is also the main cause for highest death (15%) among all types of cancer, distinctively, in women. This trend can be reviewed in Malaysia, as well. According to the Malaysian Cancer Society, it is the number one cancer that brings death to women (around 25% of the total number), in addition it is the commonest cancer among women. Based on the calculation, among 100 women in Malaysia, now, about 5 will have breast cancer. Meanwhile, for those women, living in Europe and US, the chance contract their breast cancer, is around 12.5%. Sample 's add to them problem, these women with breast cancer in Malaysia often come to hospital very late, unlike women in developed countries, so much emphasizes the need of early detection and diagnosis. While breast cancer although has some features of symptoms, but sometimes there is no noticeable symptoms, only by regular breast cancer screening for early discovery. Early diagnosis is critical in helping to ensure the best hope for survival, and diagnosis, the more likely the progression of a tumor and the more problems getting it under control. Much research has shown that starting therapy within three months of the onset of the first symptoms means survival is better than if therapy is delayed. For example, a systematic review by Prof MA Richards and colleagues found compelling evidence for improved survival if therapy is started within three months of the onset of symptoms, compared with delayed initiation. Realizing the need to such advancements specially in arena of machine learning, this paper provides an overview on the mechanism and

functionalities of machine learning techniques, so as to carry out breast cancer detection earlier. The intelligence mechanism of the human beings can be deployed using Artificial Intelligence (AI) applications which in turn has power integrated with Machine Learning (ML) methods to predict and aid in early detection and to take sound decision of Breast cancer detection and diagnosis.

2. LITERATURE REVIEW

1. In this section, some of the related works previously done on breast cancer diagnosis by researchers using different machine learning approaches are discussed.
2. An article on early detection of breast cancer through to the utilization of SVM classifier technique was presented by Y.Iraneus Anna Rejani and Dr.S.Thamarai selvi [12] 2009. In this piece, authors have shown how tumors can be identified from mammograms. For this purpose they specified an algorithm for tumor detection; their proposed method involves using mammograms images that have been filtered using Gaussian filter with standard deviation based on matrix dimensions like number of rows or columns.
3. Muhammet Fatih Ak [9] used the dataset of Dr. William H. Walberg from the University of Wisconsin Hospital. In this data set, visualization and machine learning techniques were applied such as logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, decision tree, random forest and rotation forest. R, Minitab and Python were selected for these machine learning techniques and visualization. All the methods have been compared with each other in a comparative analysis. The highest classification accuracy (98.1%) was obtained with the logistic regression model with all features included in our results; also, it has been shown that there is improvement on accurate performance based on our approach.
4. Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications" (2020).
5. A comparative analysis was conducted by Dana Bazazeh and Raed Shubair [1] on three machine learning techniques: Support Vector Machine (SVM), Random Forest (RF), and Bayesian Networks (BN). For the training set, they employed the Wisconsin original breast cancer dataset. Based on selected methods, classification performance is proven to differ in simulation results. It has been shown that SVMs are the most accurate about sensitivity but also specificity as well as precision too while RFs have highest probability of correct tumor classification.
6. For classifying the Diabetic disease dataset, Ou et al. [4] compared naïve Bayes, decision tree and random tree to identify which among these algorithms achieve better results. It was revealed by this research that the most appropriate classifier is naïve Bayes with 76.3% accuracy rate.
7. To discover the optimal approach for breast cancer predictions, Wang et al. [6] performed data mining techniques on multiple records. They employed support vector machine (SVM), artificial neural network (ANN), naïve Bayes classifier as well as AdaBoost Tree.
8. Williams et al. [9] made studies about risk prediction on breast cancer by using data mining classification techniques. In Nigeria, breast cancer is the most widespread form of cancer among women. There are few facilities that can predict this type of disease early enough to help. For that reason, they required a more effective approach to anticipate it as indicated by Health 2020;8(2):111-123. Two data mining methods used were naïve Bayes and J48 decision trees.

3. PROPOSED WORK

DATA SET

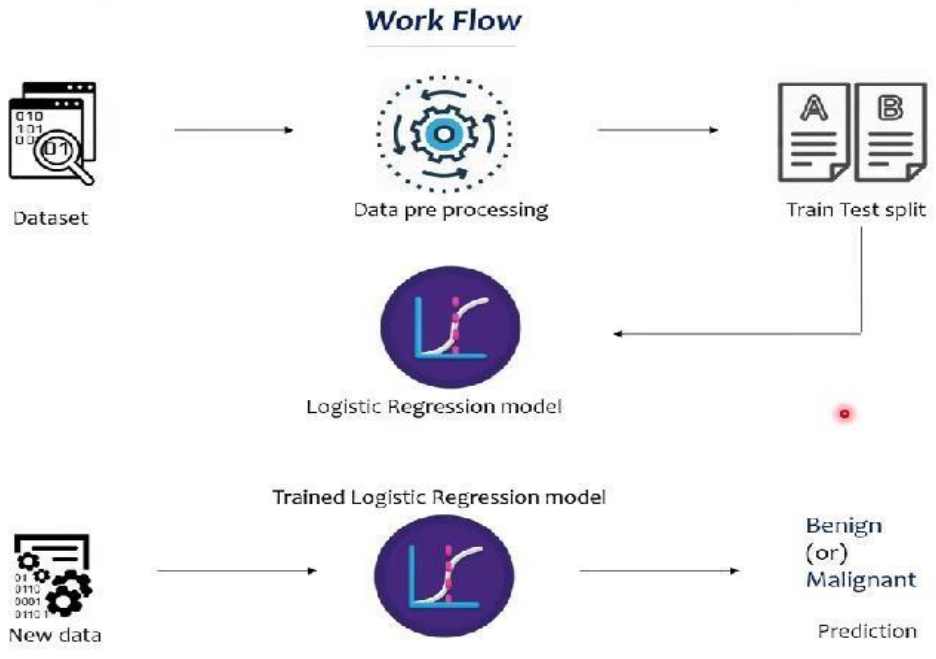
From Kaggle, the data used in conducting these various experiments was obtained. In total, this dataset consists of eleven directories corresponding to attributes like diagnosis, radius_mean, texture_mean, area_mean etc. This dataset has a total of 7,858 instances which are distributed in these eleven magnification columns. Each magnification directory consists of two folders representing tumours which are Benign and Malignant.

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
0	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
0	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
0	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
1	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766
1	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811

```
     mean radius  mean texture  mean perimeter  mean area  mean smoothness  \
0          17.99         10.38          122.80     1001.0          0.11840
1          20.57         17.77          132.90     1326.0          0.08474
2          19.69         21.25          130.00     1203.0          0.10960
3          11.42         20.38           77.58      386.1          0.14250
4          20.29         14.34          135.10     1297.0          0.10030
..           ...           ...             ...        ...              ...
564        21.56         22.39          142.00     1479.0          0.11100
565        20.13         28.25          131.20     1261.0          0.09780
566        16.60         28.08          108.30      858.1          0.08455
567        20.60         29.33          140.10     1265.0          0.11780
568         7.76         24.54           47.92      181.0          0.05263

     mean compactness  mean concavity  mean concave points  mean symmetry  \
0             0.27760         0.30010              0.14710         0.2419
1             0.07864         0.08690              0.07017         0.1812
2             0.15990         0.19740              0.12790         0.2069
3             0.28390         0.24140              0.10520         0.2597
4             0.13280         0.19800              0.10430         0.1809
..                ...             ...                  ...            ...
564           0.11590         0.24390              0.13890         0.1726
565           0.10340         0.14400              0.09791         0.1752
566           0.10230         0.09251              0.05302         0.1590
567           0.27700         0.35140              0.15200         0.2397
568           0.04362         0.00000              0.00000         0.1587
```

PREPROCESSING



Feature Selection

The importance of feature selection is inevitable in machi ne learning models. It clarifies information and reduces the complexity of information. It also reduces the data size, makes it easier to train the model and shortens the training time. It prevents data from overfitting. Choosing the best feature subset from all features can improve accuracy. Some options include wrapping method, filter method, and e mbedding method.

Recursive Feature Elimination (RFE)

The algorithm is referred to as a wrapper feature selection algorithm. It provides and uses multiple machine learning algorithms wrapped by RFE at the root of the path for selecting features. It works in such a way that each feature is scored by the filter, which selects the feature with the largest score (or smallest). Technically speaking, internally RFE uses filter based feature selection as well as being a wrapper feature selection algorithm. What happens is starting with all features in the training data; it looks for a specific subset and then removes features until n number desired is achieved. On top of model overlaying machine learning algorithms rank most important features, discard most important ones and retune model. This process continues until storage of several features takes place.

Segmentation

The image is divided into blocks of 2X2, 3x3 and up to 10 X10, which we call segmentation. During this segmentation process, we train the system to identify adjacent regions of interest that are important for BC diagnosis. By removing irrelevant information from images, tumors can be easily detected at an early stage. Kmeans clustering algorithm is a grouping method, meaning that similar objects are grouped together in the same group. The segmentation process relies on this to get better results and gives better results when there are similar products in a group. It can be processed faster than broken data [1].

MACHINE LEARNING ALGORITHMS

Support Vector Machine (SVM)

Think of the Support Vector Machine (SVM) algorithm as a tool. It tries to find something called a hyperplane in a certain space. This space has as many dimensions as there are features. The job of this tool? It separates different types of data points. During this, it can spot several hyperplanes. But the main aim is to locate a hyperplane with the widest gap, or 'margin' between the different types of data points. What's a hyperplane for? It marks the boundary for sorting the data points. Depending on where the data points are in relation to the hyperplane, they're tagged as this or that type.

SVM is a great way to classify things. It's best when there's a clear separation and lots of items in the data. But, if the dataset is too big, it can take too long to train. When there-'s a lot of noise in the data, SVM might not do as well. But even with these issues, SVM is great for classifying things in the right settings.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor or KNN is a simple Machine Learning method. It's based on Supervise-d Learning ideas. What it does is, it looks for similarities between new and old data, then groups the new data with the- most matching category. It keeps all the- usable data and groups a new piece- based on how much it matches the old data. How does KNN work? It finds data points in the set that look nearly like- the fresh point given to the- machine. The algorithm then sorts these similar points, using their distance from the- new point. We usually use Euclide-an distance for this. After that, a certain number of close

points get classified into separate groups. It's important to note that KNN often uses odd numbers of points. This keeps a balance- even when we have two classes.

The KNN algorithm is simple- and can handle big data well. However, it takes a lot of computer power because it measures distances between all data points in the- training set. Finding the best K value- also makes the algorithm more complex.

KNN is a supervised learning method, meaning we already know the- data labels before making any pre-dictions. It's useful for clustering and regression jobs. The numerical K value stands for the- closest neighbors. Unlike some- other algorithms, KNN doesn't have a separate training phase. Instead, it predicts using the Euclidean distance to the- k-nearest neighbors.

When we use KNN on a breast cancer dataset, it excels because it already has labeled data like "malignant" or "benign". The algorithm gives a label based on how close the- new data point is to the known labels in its neighborhood. The attached figure illustrates how KNN can be simple yet effective in making predictions for datasets with labels.

Random Forest

Think of Random Forest as a big team of Decision Trees. Each tree in this team is like an investigator who is making decisions. It keeps asking questions about specific data features. Then, it guides the decisions to leaf nodes. These node-s are like the answers or results. To build these trees, we use methods like Recursive Partitioning or Conditional Inference Tree-.

With Recursive Partitioning, it’s a step-by-step game. Every node, or que-stion, is a decision point. We learn by breaking the data into parts using attribute value checks. It continues until the answers are- all the same. The other method, the Conditional Inference Tree, is more- technical. It uses special tests to decide how to split the data, helping us to avoid overfitting by correcting for multiple testing.

Random Forest works well with lots of data, both small and big, and of different types. But there's a limit because too much data can cause memory issues. For big data, we need to adjust the- hyperparameters. It's like- fine-tuning, it helps prevent overfitting.

The ensemble structure of a Random Forest model, comprising many Decisions Trees, provides benefits over individual decision trees in certain scenarios, particularly at the point of regularization where model quality is optimized. Variance- and bias issues are balanced by constructing multiple- Decision Trees using random sample-s drawn with replacement. This technique helps Random Forest overcome limitations inherent to single- Decision Trees. Each tree predicts observations separately, and the ultimate decision is determined by a majority vote-. Random Forest also proves useful in an unsupervised role for assessing similarities among data points. This flexibility combined with its capacity to handle complex datasets while minimizing overfitting solidifies Random Forest as a robust and adaptable algorithm applicable across machine- learning problems.

Logistic Regression

In the realm of predictive modeling, logistic regression steps in where linear regression falls short, especially when dealing with categorical data. Unlike linear regression, which is geared towards predicting continuous variables, logistic regression is tailored for predicting binary outcomes – essentially determining the probability of something being true or false. This makes it a powerful tool for classification tasks. The logistic regression model employs the sigmoid function to transform independent variables into a probability expression ranging from 0 to 1 concerning the dependent variable. This ability to provide probabilities and classify new samples using both continuous and discrete measurements renders logistic regression a widely used machine learning algorithm. Logistic regression was initially

employed in biological studies during the early twenties. However, now it has become widespread and is applicable to social sciences and prediction cases. It is useful when one of the dependent variables is binary and others are independent of each other, particularly, since linear regression is better for continuous variables, but logistic regression works in case where the outcome is categorical. Forward propagation as well as backward propagation represent the key stages of logistic regression process. In forward propagation, the weights are used to multiply with features while sigmoid function generates probabilities based on these weights. The initial estimate then compared against observed values thus a loss function that measures the difference between predicted and real values generated. If there appears to be large gaps from the loss function, one gets into backpropagation process which aims at refining model predictive accuracy by updating weight values using derivative of cost function.

Logistic Regression Model training

```
Logistic Regression

[21] model=LogisticRegression()

#training the log reg model uing training data
model.fit(X_train, Y_train)

/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: Conve
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
LogisticRegression
LogisticRegression()
```

Logistic Regression Model accuracy evaluation

```
[23] #Model Evaluation
     #Accuracy score
     #accuracy on training data

     X_train_prediction=model.predict(X_train)
     training_data_accuracy=accuracy_score(Y_train, X_train_prediction)

[24] print('Accuracy on training data=', training_data_accuracy)

     Accuracy on training data= 0.9472527472527472

[25] #accuracy on test data

     X_test_prediction=model.predict(X_test)
     test_data_accuracy=accuracy_score(Y_test, X_test_prediction)

     print('Accuracy on test data=', test_data_accuracy)

     Accuracy on test data= 0.9298245614035088
```

Naïve Bayes

The Naïve Bayes classifier is an example of a supervised learning algorithm created for classification tasks. It is rooted in the Bayes theorem which follows a probabilistic approach to establish the likelihood of an event that happens after another one.

Naïve Bayes may seem too simple, but it is a powerful machine learning algorithm used in many industries.

The fundamental assumption underlying Naïve Bayes is that all predictors or features are independent, although this independence is rarely observed in real life. This drawback makes the model less useful in real-world situations. One such limitation involves the problem of zero frequency where it assigns zero probability to a categorical variable does not present in the training dataset but appearing in the test dataset. This can be handled by applying smoothing methods.

To achieve optimal accuracy, Naïve Bayes depends on large datasets, which presents a major challenge because acquiring broad dataset cannot always be possible. Nonetheless, Naïve Bayes still remains one of the most effective classifiers; it has been applied successfully to various classification problems through probabilistic reasoning.

PROPOSED METHODOLOGY

The primary objective of our project is to utilize logistic regression to classify tumors as either malignant or benign, subsequently analyzing the tumors. The results obtained will provide insights into the practical application of the model.

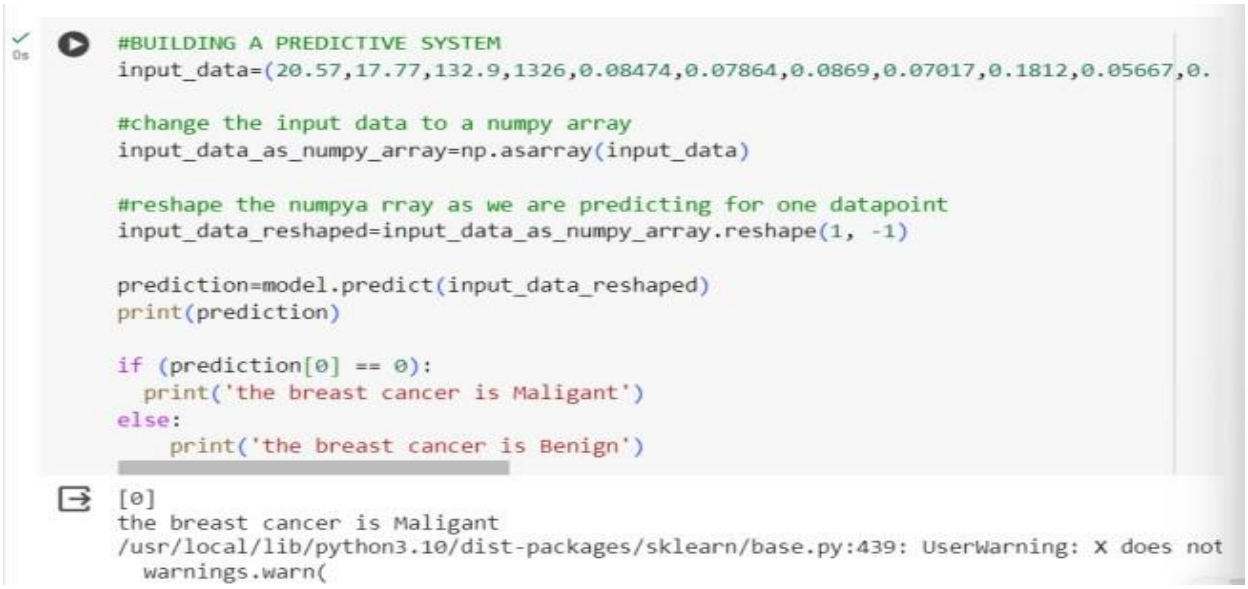
1. To design this logistic regression model, the following steps will be followed:
2. Import all necessary libraries for data processing and logistic regression modeling.
3. Create a dataset containing images of tumors and their corresponding labels (malignant or benign).
4. Preprocess the image data and associate each image with its corresponding label.
5. Train the LR model on the image dataset to learn the distinguishing features of malignant and benign tumors.
6. After this step, the data needs to be split into testing and training data.
7. Implement logistic regression as the classification algorithm, utilizing the pixel values of the tumor images as input features.
8. Evaluate the performance of the logistic regression model using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score.

In logistic regression for image recognition:

- Each pixel in the tumor image serves as a feature input to the logistic regression model.
- The logistic regression model applies a linear transformation to the input features, followed by the logistic sigmoid function to produce the predicted probability of the tumor being malignant or benign.
- Through iterative optimization algorithms like gradient descent, the logistic regression model learns the optimal weights for each feature to minimize the classification error.
- The final output of the logistic regression model is a probability score indicating the likelihood of the tumor being malignant or benign.

By employing logistic regression for breast cancer detection, we aim to provide a streamlined and interpretable approach for distinguishing between malignant and benign tumors, contributing to early detection and improved patient outcomes.

Results



```
#BUILDING A PREDICTIVE SYSTEM
input_data=(20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.

#change the input data to a numpy array
input_data_as_numpy_array=np.asarray(input_data)

#reshape the numpya rray as we are predicting for one datapoint
input_data_reshaped=input_data_as_numpy_array.reshape(1, -1)

prediction=model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
  print('the breast cancer is Maligant')
else:
    print('the breast cancer is Benign')
```

```
[0]
the breast cancer is Maligant
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not
  warnings.warn(
```

Conclusion

In this article, we review different study methods for the diagnosis of breast cancer. We made a comparison between Logistic Regression, CNN, KNN, SVM, Naive Bayes and Random Forest. It turns out that Logistic Regression outperforms existing methods in terms of accuracy, precision, and dataset size.

References

1. Dana Bazazeh and Raed Shubair "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" (2016).
2. David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu “Machine Learning Classification Techniques for Breast Cancer Diagnosis” (2019).
3. Kalyani Wadkar, Prashant Pathak and Nikhil Wagh “Breast Cancer Detection Using ANN Network and Performance Analysis with SVM” (2019).
4. Qu, Z. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2011, 9, 515.
5. SH Nallamala, P Mishra, SV Koneru - Int J Recent Technol Eng, 2019 - academia.edu
6. Wang, H.; Yoon, W.S. Breast cancer prediction using data mining method. In Proceedings of the 2015 Industrial and Systems Engineering Research
7. Conference, Nashville, TN, USA, 30 May–2 June 2015
8. Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications" (2020).
9. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning" (2019).
10. Williams, T.G.S.; Cubiella, J.; Griffin, S.J. Risk prediction models for colorectal cancer in people with symptoms: A systematic review. BMC Gastroenterol. 2016, 16, 63.
11. Mohammed, Siham A., et al. "Analysis of breast cancer detection using different machine learning techniques." *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5*. Springer Singapore, 2020.

12. Priyanka, Kumar Sanjeev. "A review paper on breast cancer detection using deep learning." *IOP conference series: materials science and engineering*. Vol. 1022. No. 1. IOP Publishing, 2021.
13. Y Rejani- "Early detection of breast cancer using SVM". 2009 -arxiv
14. Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong "Histopathological breast-image classification with restricted Boltzmann machine along with backpropagation." (2018).
15. Chauhan, Alok, et al. "Breast cancer detection and prediction using machine learning." *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2021.
16. Sebastien Jean Mambou , Petra Maresova , Ondrej Krejcar , Ali Selamat and Kamil Kuca.
17. "Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model" (2018).
18. Al Bataineh, Ali. "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection." *International Journal of Machine Learning and Computing* 9.3 (2019): 248-254.
19. M. Tahmooresi , A. Afshar, B. Bashari Rad , K. B. Nowshath and M. A. Bamiah "Early Detection of Breast Cancer Using Machine Learning Techniques".
20. Bayrak, Ebru Aydındag, Pınar Kırcı, and Tolga Ensari. "Comparison of machine learning methods for breast cancer diagnosis." *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT)*. Ieee, 2019.
21. Naji, Mohammed Amine, et al. "Machine learning algorithms for breast cancer prediction and diagnosis." *Procedia Computer Science* 191 (2021): 487-492.
22. Hussain, Lal, et al. "Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies." *2018 17th*