International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 03 | Mar 2023 www.irjet.net p-ISSN: 2395-0072

Breast Cancer Prediction using Machine Learning

Himon Sarkar[1]

[1]Student, Department of Computer Science and Engineering, Heritage Institute of Technology ,West Bengal, India.

Abstract - Breast cancer is a significant public health issue, affecting millions of women worldwide. Worldwide, female breast cancer has now surpassed lung cancer as the most commonly diagnosed cancer. An estimated 2,261,419 new cases of breast cancer were diagnosed in women across the world in 2020.. Breast cancer tumors can be classified as either benign or malignant. Fibroadenomas are solid, smooth, firm, noncancerous (benign) lumps and they do not spread to other parts of body. They may cause discomfort or pain, but they are not life-threatening. On the other hand, ductal carcinoma in situ, invasive ductal carcinoma, inflammatory breast cancer, and metastatic are malignant tumours, which are cancerous growths that have the potential to leave the breast tissue and invade other bodily organs. The distinction between benign and malignant breast tumors is important because it helps determine the appropriate course of treatment. Here Machine Learning (ML) comes to tremendous help as it can accurately predict the type of tumor by analysing large amounts of data and intricately complex patterns. In this paper, we have identified this problem as a Binary Classification problem and have implemented four different classification techniques, namely logistic regression, support vector machines, random forests, and neural network, to predict breast cancer based on patient data and imaging results. The algorithms' accuracy results were carefully studied and it was found that Logistic Regression gave the highest accuracy rate, reaching up-to 98.24%.

***Key Words*: Machine Learning, Breast Cancer, Binary Classification, Logistic Regression, Random Forest Classifier, Support Vector Machines, Artificial Neural Network, Cancer Dataset, Malignant & Benign Tumors**

1.INTRODUCTION

The accurate diagnosis of some crucial information is a major problem in the area of bioinformatics or medical science. In the field of medicine, illness diagnosis is a challenging and labor-intensive task. Numerous diagnostic facilities, hospitals, and research facilities, as well as a number of websites, all have access to a vast quantity of medical diagnosis data. It is hardly ever essential to classify them in order to automate and speed up disease diagnosis. The medical planning officer's knowledge and expertise in the medical area are typically the foundation for the disease diagnosis. This leads to situations where errors and unintended biases occur, and it takes a long time to accurately diagnose an illness.

Breast cancer is a disease that occurs most frequently. Over three million women are thought to be impacted yearly. Five-year survival for breast cancer among women diagnosed between 2015 and 2020 varies greatly with variations in location. In most places, it is widely known to be greater than fifty percent. Although there is no known way to prevent breast cancer, odds of survival are greatly increased by early detection and diagnosis.

Early in the course of the illness, the symptoms are not well-presented, which delays identification. The National Breast Cancer Fund (NBCF) advises that women over the age of forty should get a mammogram once a year. An X-ray of the breast is what a mammography is. It is a medical technique used to find breast cancer in female patients without causing any negative side effects, making the process safe. Women who undergo mammograms have a higher chance of survival than women who do not.

For characteristic tumors, automation of the identification method is therefore essential. Many people have already attempted using machine learning techniques to identify cancers in their family members, and researchers have also verified that these algorithms are more effective at doing so. The application of machine learning algorithms on breast cancer in women is summarised in this article. A malignant tumour is one that develops and spreads throughout the body. Therefore it's necessary to comprehend priorly. Thickening or having a lump on breast is considered of as symptoms of carcinoma. The main goal of the paper is to categorise and determine whether or not an individual has malignant tumors.

2. PRIOR WORKS

[1] This paper is a comprehensive review of the different machine learning algorithms used for breast cancer prediction. The authors compare the performance of different algorithms, including SVM, artificial neural networks, decision trees, and logistic regression. They also discuss the challenges and limitations of these algorithms

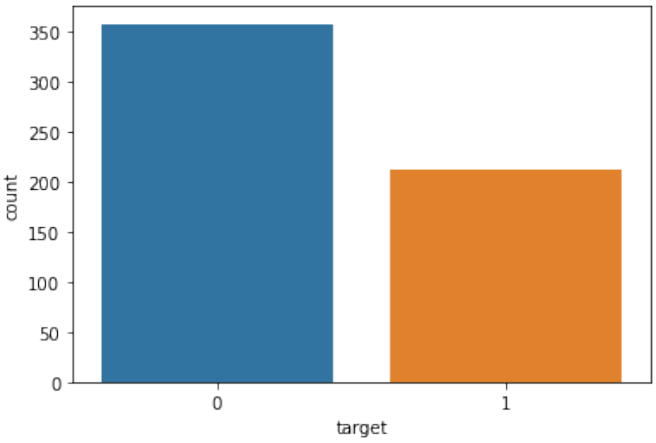
and suggest future research directions. The authors found that SVM achieved the highest accuracy in most studies, followed by artificial neural networks and decision trees. They also noted that the choice of features and data preprocessing techniques have a significant impact on the performance of these algorithms. The review concludes with a summary of the key findings, highlighting the potential of machine learning algorithms in improving the accuracy of breast cancer diagnosis. The authors emphasise the need for further research to address the challenges associated with using machine learning.

[2] In this study, a breast cancer detection method based on convolutional neural networks (CNN) is proposed. To teach their CNN algorithm, the authors used a dataset of mammogram images. They came to the conclusion that CNN-based models can be a helpful instrument for diagnosing breast cancer after achieving an accuracy of 96.38% in their experiments. This paper's use of a sizable collection of mammogram images to train the CNN model is one of its strong points. The performance of the CNN model outperformed other ML algorithms, such as SVM and logistic regression, according to the writers' comparisons. The dataset's possible biases and the CNN model's poor interpretability are just two examples of the study's limitations that the authors failed to address. The results also show that the proposed CNN-based model outperforms other approaches, achieving an accuracy of 97.18%, sensitivity of 97.47%, and specificity of 96.60%. The authors also provide a detailed analysis of the model's performance, highlighting its ability to correctly identify cancerous lesions.

[3] The authors of this study compared various feature selection and classification approaches for machine learning-based breast cancer prediction. They evaluated the effectiveness of various feature selection techniques, such as principal component analysis and mutual information, and various classification algorithms, such as SVM and decision trees, using a dataset of mammogram images and patient data. In their experiments, the authors discovered that the SVM algorithm with mutual information feature selection had the best accuracy. They also talked about how crucial feature choice is to enhancing the accuracy of machine learning algorithms for predicting breast cancer. The authors suggest that the use of machine learning algorithms, particularly the combination of PCA and support vector machines, can improve the accuracy and efficiency of breast cancer prediction, leading to earlier detection and improved outcomes for patients.

3. DATASET ACQUISITION

In this paper, we have used the University of Wisconsin Hospitals Madison Breast Cancer Database's dataset for our research. A digitised image of a breast cancer sample acquired through fine-needle aspiration is used to calculate the features of the dataset (FNA). These traits allow us to infer the characteristics of the cell nuclei visible in the image. Breast Cancer Wisconsin Diagnostic has 569 instances (Benign: 357 Malignant: 212), 2 classes (62.74% benign and 37.26% malignant), and 11 integer-valued characteristics (-Id -Diagnosis -Radius - Texture -Area -Perimeter -Smoothness -Compactness -Concavity -Concave points -Symmetry -Fractal dimension).



A brief description of the above features of the dataset is given below:

FEATURES	DESCRIPTION
Radius	It is the mean of distances from the centre to the points on the circumference
Texture	The standard deviation of the grey-scale values
Perimeter	Circumference of Tumour
Area	Area of the Tumour
Smoothness	It is the local deviation in radius
Compactness	Defined as [(perimeter^2)/area - 1]
Concavity	The gravity of concave portions on the silhouette
Concave points	Number of concave portions on the silhouette
Symmetry	A balanced and proportionate similarity that is found in two halves of an object
Fractal dimension	It is a characteristic parameter used to describe the irregular extent of coastline

Table -1: Brief description about data features

4.1 EXPLORATORY DATA ANALYSIS

For our dataset, it was found that the data was relatively consistent and there were no null values in each feature column. Initially 'M' and 'B' labels were used to specify malignant and benign tumors. We have subsequently changed the labels to 1 and 0 for better representation.

We have also computed the mean value of the feature list for the malignant and benign tumors. The mean of few features are given below:

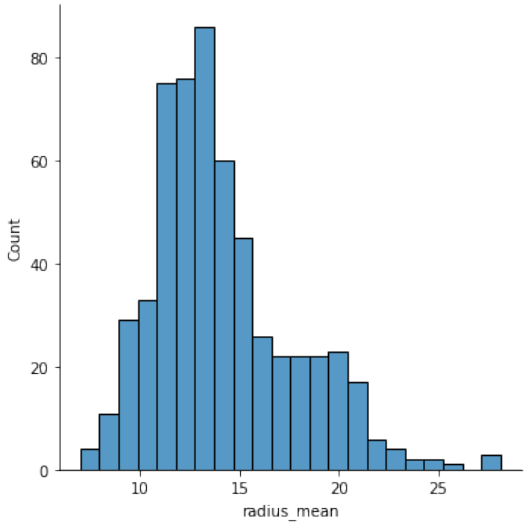
Mean Value	Target 0	Target 1
Radius	12.146524	17.462830
Texture	17.914762	21.604906
Perimeter	78.075406	115.365377
Area	462.790196	978.376415
Smoothness	0.092478	0.102898
Compactness	0.080085	0.145188
Concavity	0.046058	0.160775
Concave points	0.025717	0.087990
Symmetry	0.174186	0.192909
Fractal dimension	0.062867	0.062680

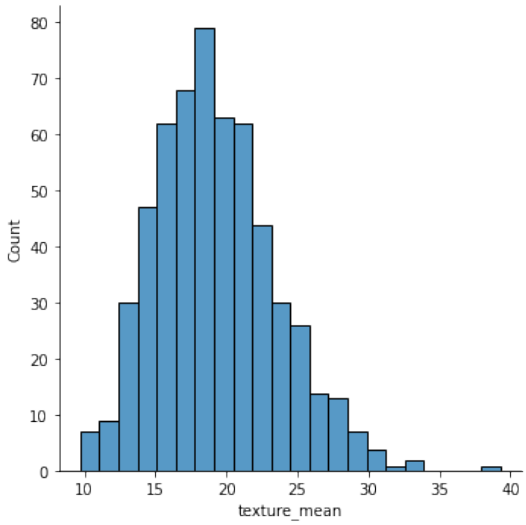
Table -2: Mean value of features for two classes

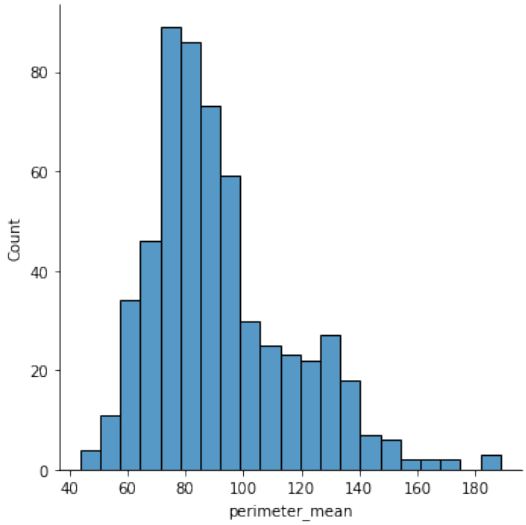
4.2 DATA VISUALISATION

Each attribute's distribution plot can have a substantial impact on how accurately the generated function works. The characteristics need to be normally distributed (Gaussian Curve or Bell Curve). The Gaussian Curve distribution plot has also helped us to identify the skewness of the data. The two main types of skewness are Positive Skewed and Negative Skewed. In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the highest value. In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.

In the following figures, the distribution plots for a few of the characteristics are displayed.







The scatter plot between the different features of the dataset has helped us to find the correlation between them. Since the scatter plot among all the features will have taken drastic time, we have show it for only first 2 features, i.e between radius_mean and texture_mean.

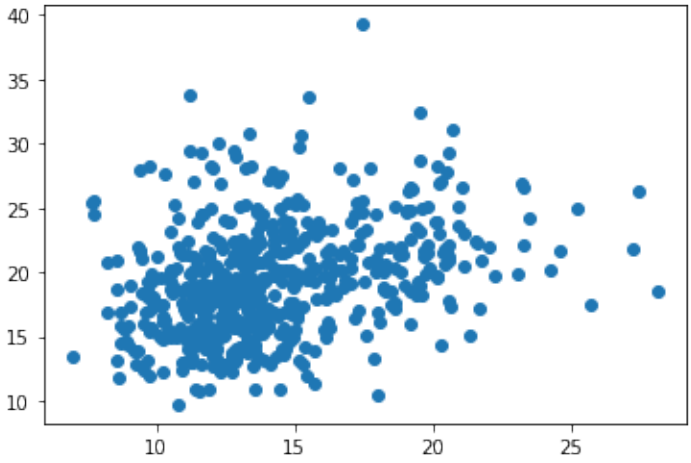


Figure 1: Scatter Plot between radius_mean and texture_mean

We have also used box plot for visualising the outliers in the dataset. Box outliers are important because they provide information about the distribution of the data. An outlier is a data point that is significantly different from other data points in the dataset. Outliers can occur due to measurement errors, data entry errors, or because the data truly represents an extreme value or an unusual event. Box outliers are typically identified using the "1.5 x IQR rule", where IQR stands for "interquartile range". The interquartile range is the distance between the first and third quartiles of the data, which defines the middle 50% of the distribution. The 1.5 x IQR rule identifies any data points that fall outside of the range of 1.5 times the IQR below the first quartile or above the third quartile. Box outliers indicate the presence of unusual or unexpected data points that may warrant further investigation. In some cases, outliers can be indicative of errors or measurement issues that need to be addressed. Box plot of radius_mean is given below:

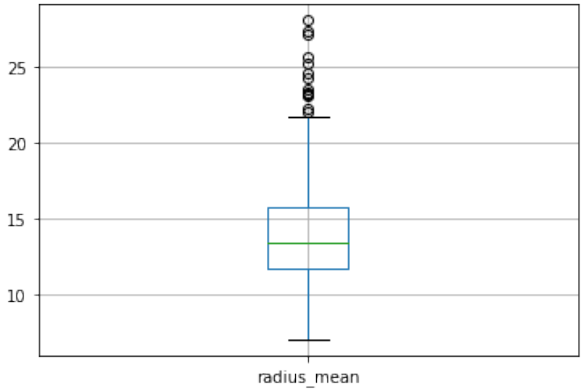


Figure 2: Outliers detection in dataset using Box Plot

The figure 3 shown below shows the heat map of the correlation between the feature groups in the dataset. The row of the heat map is represented by the value of the first component, and the column by the value of the second. The two discrete variables are now connected in a two-dimensional correlation matrix.

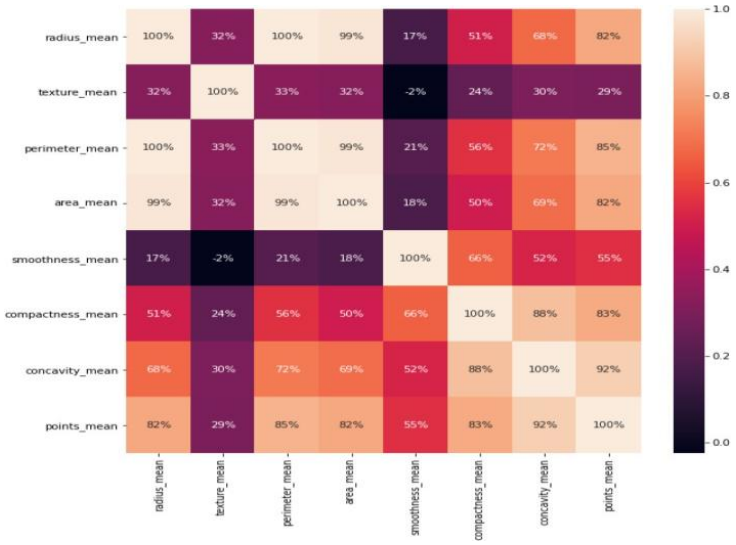
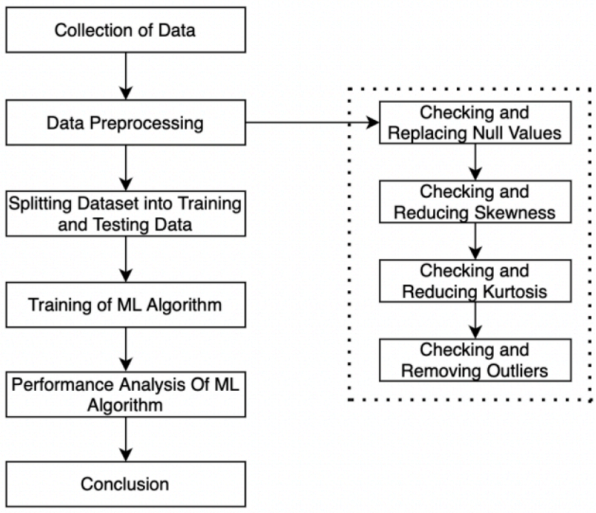


Figure 3: Heat Map of the dataset

5. METHODOLOGY

Here we have summarised the steps that we have followed in the form of a flow chart for better visualisation and understanding.



6 MACHINE LEARNING ALGORITHMS

The three classes of ML algorithms that we can generally divide into are Supervised Learning, Unsupervised Learning, and Reinforcement Learning. In this article, supervised learning algorithms are the main topic. In the instance of supervised learning, labelled input data are given to the algorithm, and it then maps the input data to well-known input-output pairs. A brief description of the algorithms that have been used here are given below:

- Logistic Regression (LR) : It is a statistical method used for analysing data in which the dependent variable is categorical. It is a type of generalised linear model that estimates the probability of an event occurring based on one or more independent variables. The output of logistic regression is a probability value that falls between 0 and 1, representing the likelihood of the event occurring.
- Random Forest Classifier (RFC) :- It is a machine learning algorithm that belongs to the family of ensemble methods. It is used for classification problems, where the target variable is a categorical variable with two or more classes. The algorithm combines multiple decision trees to make predictions and is particularly useful when dealing with high-dimensional datasets with complex relationships between the input features and the target variable. In random forest classifier, multiple decision trees are constructed using bootstrapped samples of the original data. Each decision tree is built by recursively partitioning the data into smaller subsets based on the values of the input features. At each node of the decision tree, the algorithm selects the input feature that provides the best split, according to some criterion such as the reduction in the Gini impurity or entropy. The process is repeated until a stopping criterion is met, such as reaching a maximum depth or having a minimum number of samples in a leaf node.
- Support Vector Machine (SVM) :- Support vector machines (SVMs) are a machine learning algorithm that can be used for classification and regression problems. SVMs are particularly useful when dealing with high-dimensional datasets, where the number of input features is much larger than the number of samples. SVMs work by finding the optimal hyperplane that separates the different classes in the data.In binary classification problems, the hyperplane that separates the two classes is the one that maximises the margin, which is the distance between the hyperplane and the closest data points of each class. The data points that are closest to the hyperplane are called support vectors, and they determine the position and orientation of the hyperplane. SVMs can also handle non-linearly separable data by mapping the input features into a higher-dimensional space, where the classes can be separated by a hyperplane.
- Artificial Neural Network (ANN) :- Neural networks, also known as artificial neural networks (ANNs), are a family of machine learning algorithms inspired by the structure and function of the human brain. Neural networks can be used for a wide range of tasks, such as classification, regression, clustering, and pattern recognition.A neural network consists of multiple layers of interconnected nodes, called neurons, which receive input signals and produce output signals. The input signals are processed by the neurons in the first layer, which pass their outputs to the neurons in the next layer, and so on, until the output layer produces the final prediction or classification. Each neuron in the network has a set of weights, which determine the strength and direction of the connections between the neurons. During training, the weights are adjusted to minimise the difference between the predicted outputs and the true outputs.

7. RESULTS

The entire dataset is split into a training set and a testing set before the machine learning methods have been applied. 85% of the dataset, or roughly 483 samples of data, is provided for machine learning training and the remaining 15%, or roughly 86 samples of data, is provided for testing, which is then used to predict the result. The effectiveness of machine learning methods are assessed and compared.

Machine Learning Algorithm	Training Accuracy	Testing Accuracy
Logistic Regression	95.22%	98.24%
Random Forest Classifier	95.38%	94.13%
Support Vector Machine	98.40%	97.28%
Artificial Neural Network	97.34%	97.19%

Thus it was concluded that Logistic Regression algorithm gave the highest test data accuracy of 98.24% among the other four algorithms which were implemented.

8. CONCLUSION

In conclusion, the use of machine learning methods in the early detection and prognosis of breast cancer has shown great promise. The studies examined here show that machine learning algorithms are capable of efficiently analysing vast amounts of patient data and making precise predictions about the probability of developing breast cancer. There are still issues that need to be resolved, such as ensuring the quantity and quality of the data used to train these models, addressing data bias, and putting in place efficient clinical processes that take machine learning predictions into account. Overall, the study

provided in this paper emphasises the need for further investigation and refinement of these techniques to increase their efficiency in clinical practise while highlighting the potential of machine learning in breast cancer prediction. These studies' findings imply that machine learning models can be practical aids in clinical decision-making, assisting medical workers in making defensible choices regarding patient care and treatment. By using these models, breast cancer may be discovered early, improving patient outcomes and potentially saving lives.

REFERENCES

[1] R. B. Mane and V. S. Prabhu, "Breast cancer diagnosis using machine learning algorithms: A review," International Journal of Intelligent Systems and Applications, vol. 13, no. 2, pp. 61-68, 2021.

[2] Anjali Sharma, Satish Kumar, and Rakesh Kumar, "Breast cancer diagnosis using convolutional neural networks," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 3, pp. 1431-1440, 2020.

[3] Wen-Hung Liao and Yi-Chun Chen, "Breast cancer prediction using machine learning algorithms: a comparison of feature selection and classification methods", International Journal of Medical Informatics, 2018, 114, 7-17)

[4] Ahmed FE, Ahmed NC, Vos PW, Bonnerup C, Atkins JN, Casey M. Artificial neural network modeling of breast cancer recurrence. Expert systems with applications. 2007 Mar 1;32(2):610-23.

[5] Al-masni MA, Saleh AI, Al-antari MA, Hussain M, Alghamdi WS. Computer-aided diagnosis of breast cancer using deep learning algorithms. Journal of healthcare engineering. 2018;2018.

[6] Amoah S, Sefa-Dedeh S. Breast cancer detection using support vector machines. World Academy of Science, Engineering and Technology. 2009 Sep;34:453-8.

[7] Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016 Jan 1;83:1064-9

[8] Chae EY, Kim HH, Cha JH, Kim HJ, Shin HJ, Kim HS. Radiomics features on mammography for the assessment of breast cancer. Scientific reports. 2020 Mar 4;10(1):1-0

[9] El-Said M, Ahmed SA, Ahmed FE. Artificial neural network modeling of breast cancer data: an overview. Applied soft computing. 2017 Aug 1;57:584-96.

[10] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Jan;542(7639):115-8.

[11] Jafari MH, Sadeghi M, Mashohor SB, Mahmud HR, Saripan MI, Abdul Nasir NA. Computer-aided detection/ diagnosis of breast cancer in mammography and ultrasound: a review. Clinical Imaging. 2016 May 1;40(3):279-92.

[12] Jafari-Koshki T, Arsang-Jang S, Mahjub H. Application of artificial neural network for prediction of breast cancer survival: a retrospective study in Iranian population. Journal of research in health sciences. 2014;14(4):275-9.

[13] Wang X, Peng Y, Lu L, Lu H, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 Jul 21 (pp. 2097-2106).

[14] Li X, Platel B, Chen H, Li H, Zhou Y, Chen Z, Cai W, Liu M. Breast cancer detection using deep convolutional neural networks and support vector machines. Neural Computing and Applications. 2018 May 1;29(5):1493-503.

[15] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 967-972.

[16] https://www.cancer.net/cancer-types/breast-cancer/ statistics