Comparison of Machine Learning techniques for Breast Cancer Detection

Abstract

Breast cancer is the deadliest cancer among women. Each year 1000's of women die because of breast cancer. The key factor contributing to this is late diagnosis. Current methods like mammogram screening, MRI, and liquid biopsy tests to detect breast cancer are not as efficient as they should be. Artificial intelligence researchers think that deep learning can be used by a radiologist to improve patient outcomes. Breast cancer disease classification is to classify its most common types which are malignant and benign. The latest stats show that overall 13% of the women found to be victim of this disease. This means that there is a 1 in 8 chance that the woman will develop this disease. Due to this level of risk rate, the mortality rate is very high.

In this project, we adopted some machine learning techniques and one of the deep learning custom designed neural network named as Artificial Neural Network. We compared the performances for classifying the breast cancer type which is either malignant or benign. The techniques we used are Random Forest, SVM, KNN, and ANN. We got the very best of accuracy up to 94% from the machine learning techniques while with the deep neural network we are achieving up to 96% accuracy easily after tuning some parameters.

INTRODUCTION

The rate of new breast cancer cases among women is increasing rapidly. Approximately 2.3 million women were diagnosed with breast cancer in 2022, and the mortality rate was 29%[14]. The primary reason behind the high mortality rate is late diagnosis. Key factors contributing to late diagnosis are lack of proper screening and radiologists, inaccuracy of mammograms, and traditional values [13]. The lack of radiologists and the inaccuracy of mammographic screening exacerbate the situation. One of the main challenges with screening mammograms and reading them is their complexity. Mammograms are very complicated and contain a lot of information that can be used to detect breast cancer. We need an expert mammographer to read the mammograms and detect breast cancer. Otherwise, a radiologist may miss cancer. There is also another method which is very popular and may prove to be very accurate i.e. liquid biopsy. However, it needs some more improvements and researchers are working on it. Therefore, there is a dire need for a better method to assist radiologists in the early detection of breast cancer [12].

Deep learning is a powerful tool for extracting useful information. But, it needs AI-ready data. Data that can be used for information extraction. Data is very important for AI models for breast cancer detection. In this project, we will apply our dataset onto the AI techniques to get the best model.

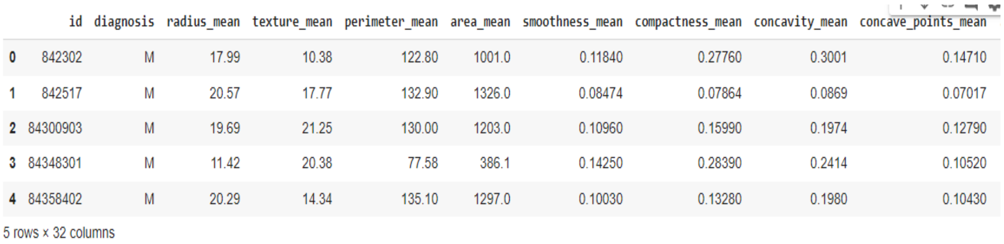
Literature Review

This section will present a wide range of literature reviews about breast cancer detection using deep learning. CNN is a widely used deep neural network architecture for image processing, image classification and other computer vision tasks. It differs from a standard artificial neural network in two ways: 1) Parameter sharing, and 2) Special convolution operation. These two properties of CNN make it computationally efficient and reliable for image processing tasks [3]. In [1], researchers have used deep neural networks for breast cancer classification. The study is conducted on the Wisconsin Breast cancer (Diagnostic) dataset. Improvement in computational efficiency, accuracy, and overfitting is achieved Fig. 1. [16] CNN Architecture. by feature selection. Irrelevant features are withdrawal using the random forest method. Researchers have concluded that with an appropriate number of hidden nodes and layers, a deep neural network can classify breast cancer with 97% to 99% accuracy. The above model will work well in some cases. However, most of the real-world data will require data preprocessing. In [2], researchers have proposed a deep learning model with hyper-parameter tuning and data pre-processing steps. This study is conducted on the WBCD dataset. The skewed data problem is solved by applying feature transformation and square root function on normalized data. Moreover, researchers used a data augmentation technique to increase the number of data. A deep neural network with a dropout of 0.5 is trained using pre-processed data. This research concluded that if we use optimized parameters like Adam optimizer and dropout of 0.5. We can achieve F1 scores of 97.44%. In a study proposed by Emmanuel, a mammogram dataset of MIAS applied for breast cancer classification. Mammograms are a structural representation of women's breasts. It is examined by a radiologist to detect lesion areas of the breast which may have cancerous cells. The objective of this research was to train a deep learning model for breast cancer classification into benign and malignant cases by using mammograms.

Data Exploration

The dataset we have picked from the uci repository named as wisconsin breast cancer which contains 32 columns and 570 training samples. 31 of the columns are features while one column 'Diagnosis' is the target variable. The features are normally showing the texture, geometrical aspects of the breast which are named as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, etc. Almost all of the features are continuous valued types.

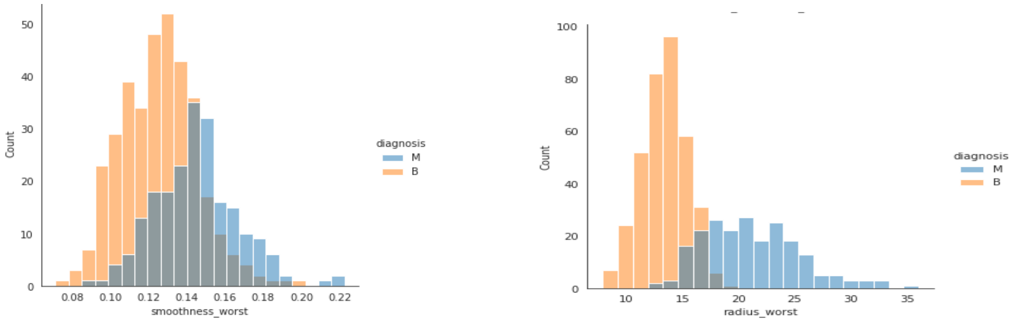
Below the samples of our dataset used:



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

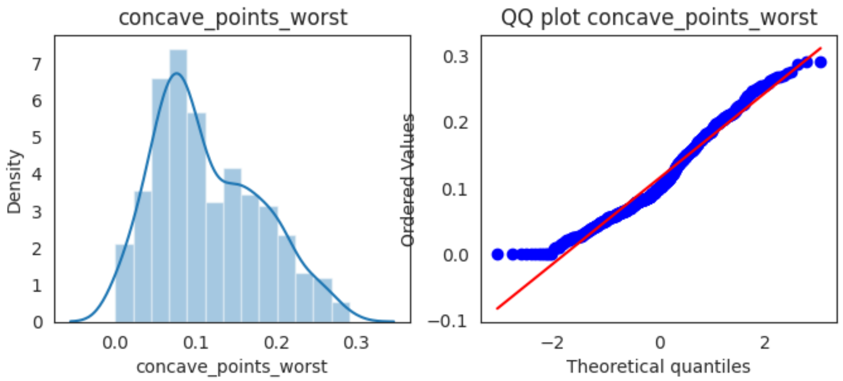
5 rows × 32 columns

Now, we will analyze and explore our dataset from a univariate, bivariate and multivariate perspective to describe it graphically. It will give us a better intuition of the data.



Above figures show the univariate analysis of the two features segmented upon target variable diagnosis. Blue bars show one type of data for one disease while orange bar plot shows the plots for the other class.

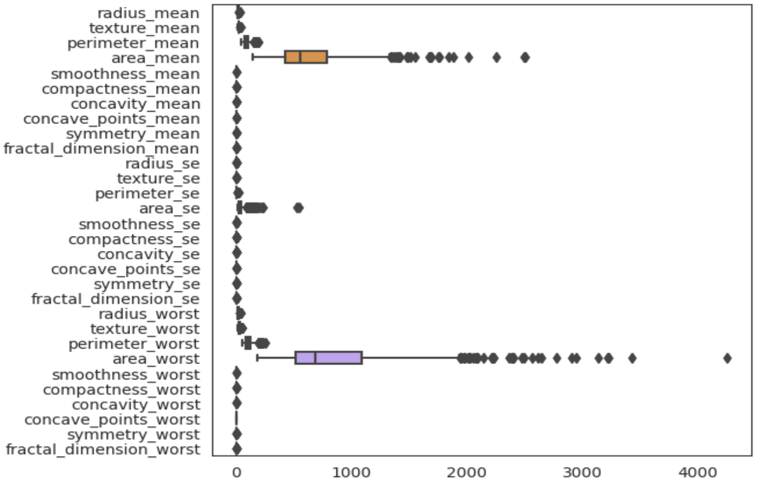
Similarly, bivariate analysis of some features is shown below in figure that either the data has features which fit linearity using QQ prob plot:



Above plot shows that the feature has skewed a bit from the left side. Hence, our feature has linear relation and the distribution is of gaussian form.

On the left side, we are observing the feature distribution and data spread is either normalized or gaussian type. While , on the right side, we are checking the linearity of the data. Below we are observing the outliers if any feature contains:

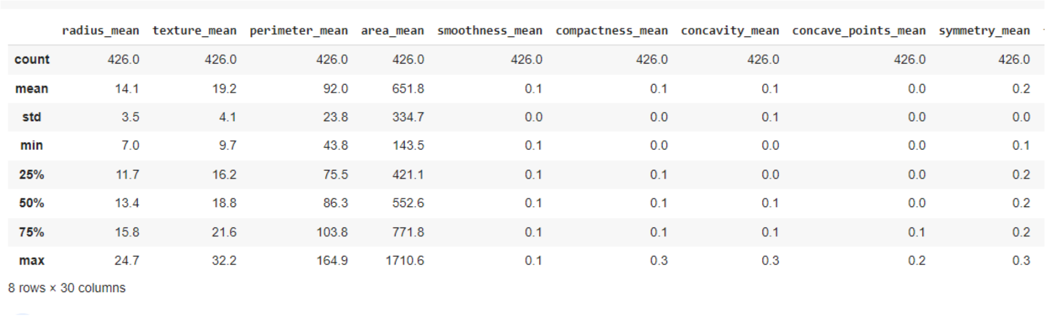
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f83cfe0b1d0>
```



Data Preparation

First, we loaded the dataset and checked its stats and info. After confirming the number of continuous valued features number and categorical features, we handled the missing values but as there are not any missing values. Hence , we head on. Then, we handled the categorical features in which the target variable ‘diagnosis’ was included and we mapped its binary values with 0 and 1. Then, we handled the numerical features in which we removed the outliers using the z-score method.

Then , we splitted the dataset into X and Y variables. After that we scaled the X features columns using the StandardScaler. Then, we splitted the X,y variables into training and testing sets for the further biased evaluation. After the whole data preparation, we have the data which would look like :



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	symmetry_mean
count	426.0	426.0	426.0	426.0	426.0	426.0	426.0	426.0	426.0
mean	14.1	19.2	92.0	651.8	0.1	0.1	0.1	0.0	0.2
std	3.5	4.1	23.8	334.7	0.0	0.0	0.1	0.0	0.0
min	7.0	9.7	43.8	143.5	0.1	0.0	0.0	0.0	0.1
25%	11.7	16.2	75.5	421.1	0.1	0.1	0.0	0.0	0.2
50%	13.4	18.8	86.3	552.6	0.1	0.1	0.1	0.0	0.2
75%	15.8	21.6	103.8	771.8	0.1	0.1	0.1	0.1	0.2
max	24.7	32.2	164.9	1710.6	0.1	0.3	0.3	0.2	0.3

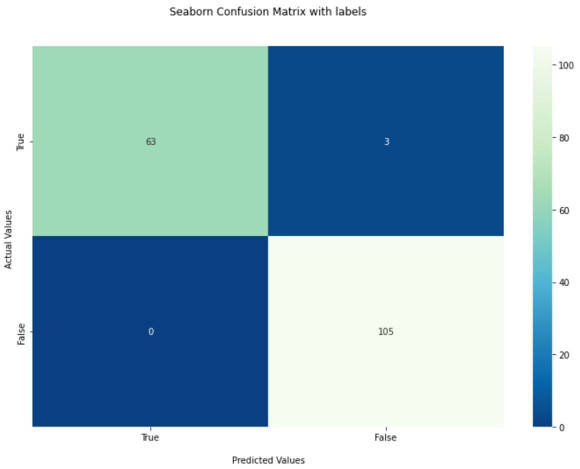
8 rows × 30 columns

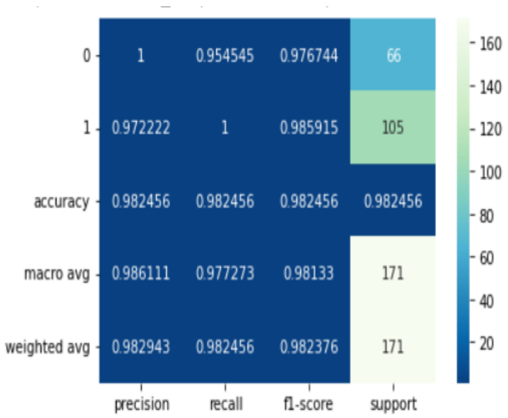
Experimental Procedure

We have used the three classification algorithms which perform best of all. These models include Logistic Regression CV, Support Vector Machines and Artificial Neural Network. First two algorithms belong to machine learning techniques and the last one is deep neural networks . logistic regression classifies the entities using sigmoid activation function with the gradient descent algorithm. While the support vector machine uses laplacian arguments to find the best possible solution. Similarly, the deep neural network uses the adam optimizer.

- **Logistic Regression CV**

The confusion matrix and classification report shows the performance of this model. Our model has given the accuracy of 98.2% on the testing set while the precision and recall score is 98.29% and 98.24% respectively.



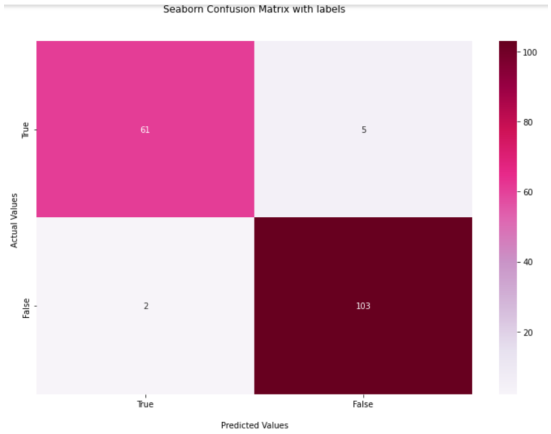


	precision	recall	f1-score	support
0	1	0.954545	0.976744	66
1	0.972222	1	0.985915	105
accuracy	0.982456	0.982456	0.982456	0.982456
macro avg	0.986111	0.977273	0.98133	171
weighted avg	0.982943	0.982456	0.982376	171

20
40
60
80
100
120
140
160

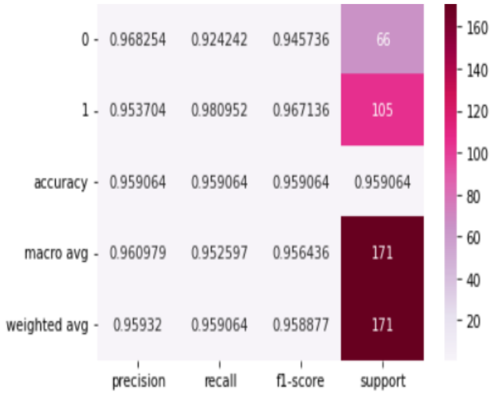
- **Support Vector Machines**

The confusion matrix and classification report shows the performance of this model. Our model has given the accuracy of 98.2% on the testing set while the precision and recall score is 98.29% and 98.24% respectively.



Seaborn Confusion Matrix with labels

Actual Values \ Predicted Values	True	False
True	61	5
False	2	103

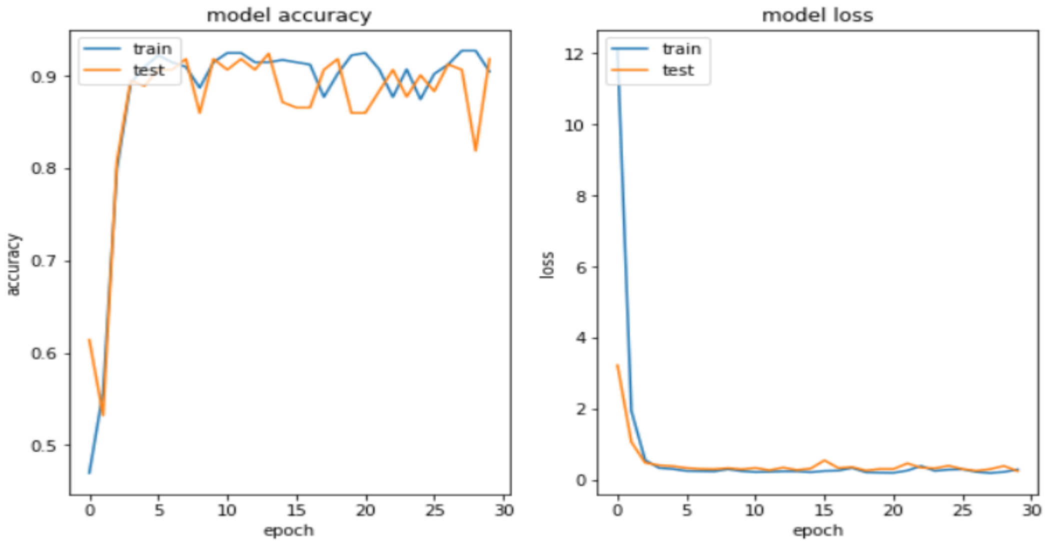


	precision	recall	f1-score	support
0	0.968254	0.924242	0.945736	66
1	0.953704	0.980952	0.967136	105
accuracy	0.959064	0.959064	0.959064	0.959064
macro avg	0.960979	0.952597	0.956436	171
weighted avg	0.95932	0.959064	0.958877	171

160
140
120
100
80
60
40
20

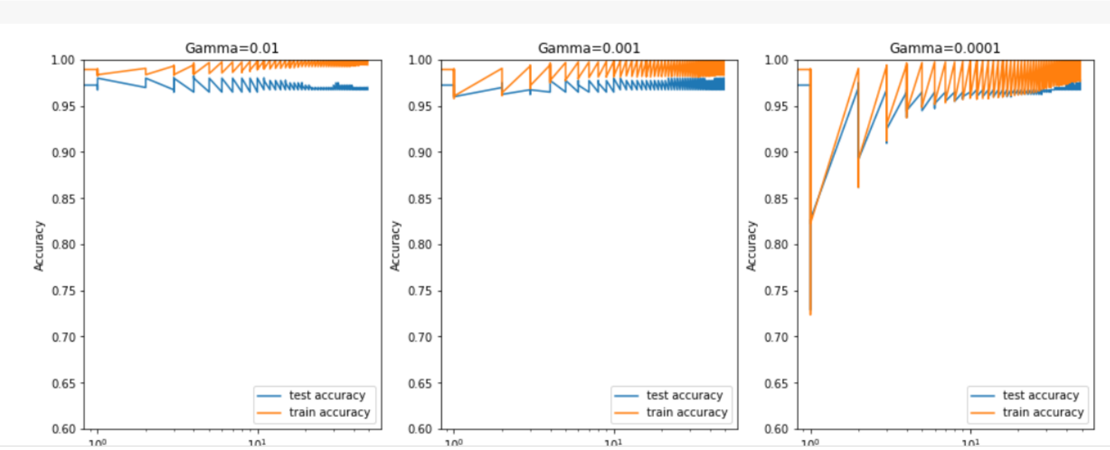
Deep Neural Network

The performance of deep learning algorithm artificial neural networks also performs much better without the tuning of the models. Below, we can see the learning curves of the model and how it avoids overfitting on the test set.



Conclusion

We tuned all the models by their random hyper parameters to see their full performance and we got that the support vector machines outperformed all the algorithms with its maximum of accuracy of 97% on testing set while on training set, it gave almost 98% of the accuracy:



References

[1] M. O. F. Goni, F. M. S. Hasnain, M. A. I. Siddique, O. Jyoti and M. H. Rahaman, ”Breast Cancer Detection using Deep Neural Network,” 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1-5, doi: 10.1109/ICCIT51783.2020.9392705.

[2] H. K. Timmana and C. Rajabhushanam, ”Breast Malignant Detection using Deep Learning Model,” 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 383-388, doi: 10.1109/ICOSEC49089.2020.9215382.

[3] E. L. Omonigho, M. David, A. Adejo and S. Aliyu, ”Breast Cancer:Tumor Detection in Mammogram Images Using Modified AlexNet Deep Convolution Neural Network,” 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), 2020, pp. 1-6, doi: 10.1109/ICMCECS47690.2020.240870.

[4] F. Yilmaz, O. Kose and A. Demir, ”Comparison of two different deep learning architectures on breast cancer,” 2019 Medical Technologies Congress (TIPTEKNO), 2019, pp. 1-4, doi: 10.1109/TIPTEKNO47231.2019.8972042.

[5] https://maelfabien.github.io/deeplearning/xception/

[6] A. Biswas, Z. Al Nazi and T. A. Abir, ”Invasive Ductal Carcinoma Detection by A Gated Recurrent Unit Network with Self Attention,” 2019 4th International Conference on Electrical Information and Communication Technology (EICT), 2019, pp. 1-6, doi: 10.1109/EICT48899.2019.9068841.

[7] D. S, M. S, K. S, S. S and M. R, ”GAN based Data Augmentation for Enhanced Tumor Classification,” 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), 2020, pp. 1-5, doi: 10.1109/ICCCSP49186.2020.9315189.

[8] S. Saranya and S. Sasikala, ”Diagnosis Using Data Mining Algorithms for Malignant Breast Cancer Cell Detection,” 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1062-1067, doi: 10.1109/ICECA49313.2020.9297481.

[9] B. Xu et al., ”Attention by Selection: A Deep Selective Attention Approach to Breast Cancer Classification,” in IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1930-1941, June 2020, doi: 10.1109/TMI.2019.2962013.

[10] S. S. Prakash and K. Visakha, ”Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks,” 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 88-92, doi: 10.1109/ICIRCA48905.2020.9183378.