Application of Machine Learning for Breast Cancer Diagnosis

A thesis submitted in partial fulfilment of the requirements

for the degree of Master of Technology

in Embedded Systems

by

Nirdosh Kumar

(2017PEB5169)

Under the supervision of

Prof. Lava Bhargava

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY, JAIPUR

June 2019

Certificate

Department of Electronics & Communication Engineering

Malaviya National Institute of Technology, Jaipur

This is to certify that the thesis work entitled " **Application of Machine Learning for Breast Cancer Diagnosis**" has been carried out by **Nirdosh Kumar** for the degree of Master of Technology in Embedded Systems at Malaviya National Institute of Technology under my supervision.

The thesis in my opinion, is worthy of consideration for award of the degree of Master of Technology (M.Tech) in accordance with the regulations of the Institute. To the best of my knowledge, the results embodied in this thesis have not been submitted to any other University or Institute for the award of any other Degree or Diploma.

Prof. Lava Bhargava

June 2019

Department of Electronics & Communication Engineering

Malaviya National Institute of Technology, Jaipur

Declaration

I declare that,

1. the thesis comprises my original work towards the degree of Master of Technology in Embedded Systems at MNIT and has not been submitted elsewhere for a degree.
2. due acknowledgement has been made in the text to all the material used.

Nirdosh Kumar
(2017peb5169)

Acknowledgements

I would like to take this opportunity to express my deep sense of gratitude and respect towards my Supervisor (Guide), **Prof. Lava Bhargava**, Professor, Department of Electronics & Communication Engineering, Malaviya National Institute of Technology, Jaipur. I am very much indebted to him for the generosity, expertise and guidance; I have received from him while working on this project and throughout my studies. Without his support, encouragement and timely guidance, the completion of my project would have seemed a far-fetched dream. He always helped me to feel motivated throughout the research work. In this respect, I find myself lucky to have him as my Project Guide. He has guided me not only with the subject matter but also taught me the proper style and techniques of working.

I would like to thank **Prof. D. Boolchandani**, HOD, Department of Electronics & Communication Engineering for his co-operation and help rendered in numerous ways for the successful completion of this work.

I take this opportunity to express my regards and obligation to my Family, whose support and encouragement, I can never forget in my life. Also, I express my gratitude to all other faculty members in the department.

I would also like to thank **Gaurav Sharma**, research scholar, for sharing the knowledge of his work and his valuable inputs.

I would also like to thanks my friend **Sachin Agnihotri**, **Rajveer Mali** and **Sangh Mitra Rathore**, Department of Electronics & Communication Engineering, for their enjoyable company, support and subject discussion during my dissertation work.

Lastly, I am thankful to all those who have supported me directly or indirectly during the dissertation work. Above all, I thank Almighty, who bestowed his blessings upon us.

Nirdosh Kumar
(2017PEB5169)

Abstract

Breast Cancer is the most prevalent form of cancer and a significant reason for high mortality rates among women. Manual diagnosis of this disease requires long hours and specialists which results in increase in the mortality rate. Therefore, an Automated breast cancer diagnosis has been proposed to reduce the time taken to diagnosis and decreases the spread of cancer. This model has been trained and tested on the Wisconsin dataset. Four different Machine Learning algorithms, namely Logistic Regression, SVM, KNN & Naive Bayes have been used for breast cancer diagnosis. The main focus of this thesis is on analyzing and comparing the accuracy, specificity & sensitivity and find the best algorithm with best accuracy machine learning (ML) algorithms, namely: logistic regression, SVM, KNN naive Bayes by calculating their classification accuracy, sensitivity specificity. The different hyper-parameters used for different ML algorithms were manually assigned. Among all the algorithms used, SVM performed best with accuracy about 98.24%

Keywords: Breast Cancer, Logistic Regression, Support Vector Machine, k-Nearest Neighbour, and Naive Bayes

Contents

List of Figures

List of Tables

Abbreviations

SVM	**S**upport **V**ector **M**achine
KNN	**K**Nearest Neighbours **S**ystem
ML	**M**achine **L**earning
WBCD	**W**isconsin **B**reast **C**ancer **Dataset**
PPV	**P**ositive **P**redictive **V**alue
NPV	**N**egative **P**redictive **V**alue
ACS	**A**merican **C**ancer **S**ociety

Symbols

C	cost of misclassification
γ	Gaussian kernel parameter
ε	strain tensor
μ	mean
λ	regularization parameter

Chapter 1

Introduction

1.1 Motivation

Breast cancer is the most common cancer in India & accounts for about 27% of all types of the cancers in women. India accounts for the 3rd highest number of cancer among women after the US and China [20]. Statistically, 1 out of 28 women is most likely to suffer from breast cancer in their life span [28]. 2000 new women are diagnosed with cancer every day and 1200 are detected at the later stages out of these 2000 women. Late detection reduces the survival rate by 3 to 17 times and costs 1.5 to 2 times higher as compared to early-stage detection. The mortality rate because of breast cancer is 1.6 to 1.7 times higher. In 2017, India had the highest mortality rates globally for breast cancer.
Early-stage diagnosis of breast cancer can help significantly in increasing the survival rate of patients [8]. Machine learning algorithms can play a vital role in the diagnosis of breast cancer [18]. This project is an attempt at reducing dependencies on the specialist for diagnosis of breast cancer.

1.2 Problem Statement

There are lots of research work going on to make human life better. But this can't be done without improving the condition of the healthcare sector. In a developing country

like India, even poor people should have access to economic and feasible medical facilities. Early detection tools of diseases will certainly be significant milestone in achieving this goal. Automization of diagnosis tools can play an important role in this thought.
After lung cancer, breast cancer is the second most fatal cancer. After analyzing the statistics of breast cancer, we figured out that India has very low number of experts for the breast cancer diagnosis as compared to the other countries, which lead to an increase in diagnosis time. This extra time taken to diagnose the disease like breast cancer can be fatal to some patients and increases the mortality rates [6]. The problem of late detection can be eradicated after automizing the diagnosis process. Use of machine learning algorithms [11] for breast cancer diagnosis is one of the examples of this idea. This will help in improving the mortality rate, reducing the dependency on the experts, decreasing the diagnosis cost and time.

1.3 Objective

The objective of this work is to implement a machine learning model which is able to diagnose breast cancer using different breast cell features such as cell radius, texture etc, with better accuracy as compared to manual diagnosis. We are using four machine learning algorithms, namely :

- Logistic regression
- Support vector machine
- k-nearest neighbours
- Naive Bayes

After implementing these four algorithms, we are comparing the test accuracy of each algorithm along with some other important parameters like PPV, NPV, sensitivity and specificity.

1.4 Literature Survey

A lot of research has been done for the automatization of breast cancer diagnosis for reducing the dependency on the experts and decreasing the diagnosis time.

Abien fred et al.[1] implemented breast cancer diagnostic model and compared the results of the six different ML algorithms, Linear Regression, SVM, Nearest Neighbor search, MLP and others on the Wisconsin Dataset. He divided the dataset into two parts: training (70%), testing (30%). He assigned the hyper-parameters for all the algorithms manually. The accuracies for both logistic regression and SVM about 96.09375%.
Shang Gao et al.[2], first sighted some issues faced in manual diagnosis techniques. Later on, he suggested the use of SVM for the diagnosis. He calculated the accuracies for different kernels and varying the parameters manually.
Al absi et al.[3] proposed an automated model for the diagnosis of breast cancer using k-nearest neighbours. It extracts various cell features of the cell obtained from the mamograms and employs KNN for the classification of the cells into cancerous or non-cancerous cells. He has used wavelet transform, statistical techniques for extracting and selecting the features.
In[4], D. Bazazeh et al. suggested that machine learning techniques can come handy for the early detection of breast cancer. He compared the results for the three algorithms, SVM, naive Bayesian and random forest for the Wisconsin dataset. He used accuracy, precision, recall and other parameters for comparative analysis.
R. Radha[9] has used k-NN in her work for breast cancer study. She proposed a model for breast cancer prognosis. She compared the results obtained from the k-NN algorithms with other clustering algorithms.

1.5 Thesis Organization

- **Chapter 1**

  The thesis starts with chapter 1 that briefs about the motivation for this work. The middle section of this chapter sheds some light on the problem statement and

objective. In the literature survey section, the research work carried out by others in the field of diagnosis of breast cancer using machine learning has been given.

- **Chapter 2**

  Chapter 2 throws some light on breast cancer and statistics related to it. This chapter further describes the different techniques used for breast cancer diagnosis such as screening or imaging tests, biopsy etc.

- **Chapter 3**

  Chapter 3 gives the details about the hardware, software, which we have used in our work. Later part of this chapter deals with the Wisconsin breast cancer dataset and its features.

- **Chapter 4**

  Chapter 4 starts with the brief introduction of supervised machine learning algorithms. Later in the chapter, all the four algorithms which we have used in our work, have been described with their mathematical model.

- **Chapter 5**

  Chapter 5 starts with the steps involved in the training the model on the Wisconsin dataset. The test results of the model are also given in this chapter.

- **Chapter 6**

  Chapter 6 covers the conclusion and the future scope along with enhancement which can be further be added to this work.

Chapter 2

Techniques for Breast Cancer Diagnosis

2.1 Breast Cancer

Breast cancer is a type of cancer that mostly occurs in women but men can also be affected by breast cancer. It originates when cells of the breast grow in an uncontrolled manner [24] (i.e. cells continue to divide, keep spreading to other tissues & don't die at the proper time) and usually transform into a tumour. This tumour becomes malignant if the cells invade the nearby tissues or spread to other areas of the human body.
It can origin from the various parts of the breast. Base on the cells where breast cancer origin, different types of breast cancer[29] are named as follow :

- Ductal carcinoma in situ(DCIS)
- Invasive ductal carcinoma (IDC)
- Invasive lobular carcinoma (ILC)
- Lobular carcinoma in situ (LCIS)
- Inflammatory breast cancer

These are the few most commonly occurring cancers in women; there are other types as well but are of least importance.

2.2 Causes of breast cancer

The factors which cause [24] breast cancer in women are obesity, hormone replacement therapy during menopause, family history of breast cancer, lack of physical exercise, long exposure to electromagnetic radiation, having children at a later age or not at all and early age at first menstruation etc.

2.3 Technique for breast cancer diagnosis

There are the following techniques for the diagnosis of breast cancer. A brief description of all these techniques is given as follows :

2.3.1 Imaging Tests[21]

- **Mammography**

  In mammography, low energy X-rays are used to examine the breast for diagnosis the abnormalities in the breast. It uses ionizing radiation for image creation. This technique is very accurate and reliable as compared to the other methods described below, but there is a higher risk of radiation exposure also. A typical mammogram [14] is shown in Figure 3.1

- **MRI**

  Magnetic resonance imaging (MRI) is one of the imaging techniques used to draw the anatomy of the different body organs. It uses strong Magnetic field for this purpose. It can detect the small breast lesions that can be sometimes are missed in mammogarms. Because of the magnetic field used in MRI, a person with ferrous implants can not be screened using MRI. Figure 4.1 is a breast MRI.

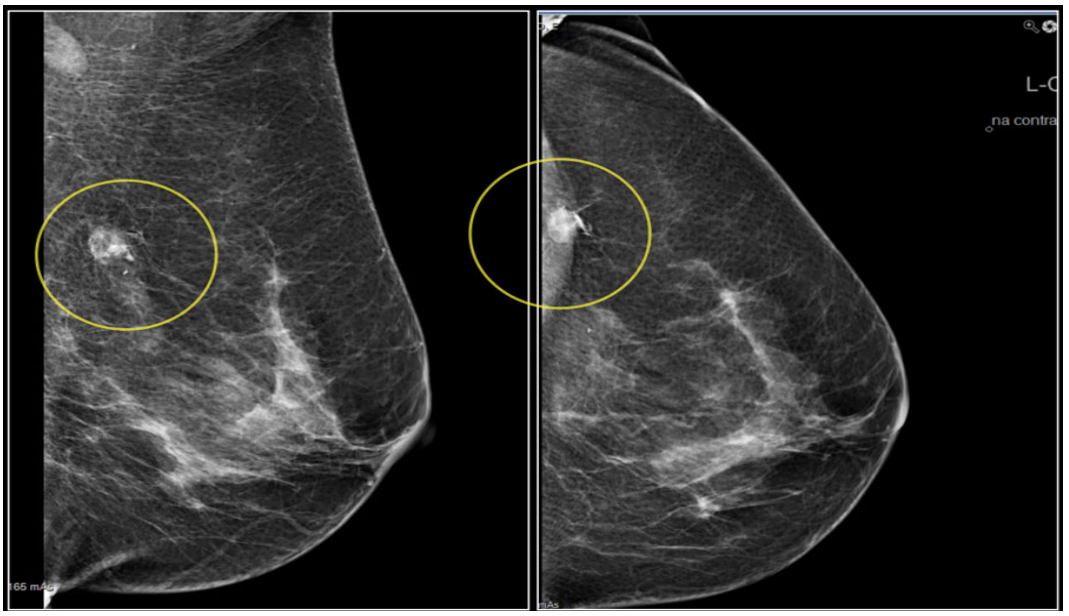

FIGURE 2.1: Mamogram[26].

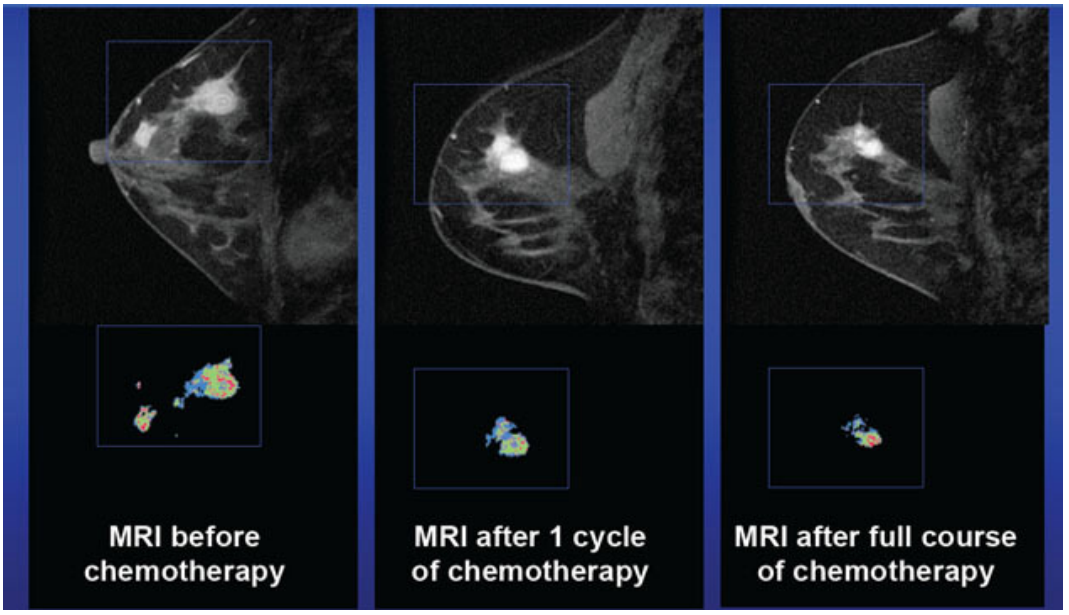

FIGURE 2.2: Breast MRI[25].

- **Ultrasound**

  It uses the ultrasound waves (frequency greater than 20,000 Hz) for creating images of the internal body structure, i.e. joints, muscles, internal organs etc. It is more useful in younger patients, because of the denser fibrous tissue of the breast which is difficult to be interpreted using mammograms. A mammogram is given in Figure

4.2.

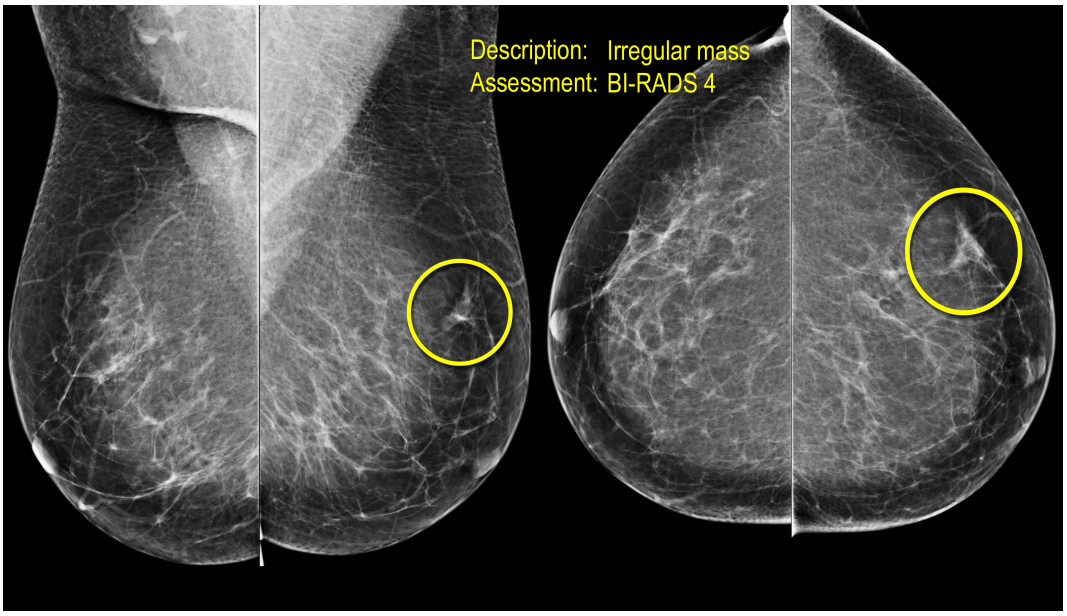

FIGURE 2.3: Breast ultrasound[26] .

All these imaging techniques characterise the breast cell present in the image using different features like cell size, the texture of cell etc. Based on these findings, we can predict whether someone has breast cancer or not.

The ACS (American cancer society) has made it mandatory to have a mammogram and and MRI for the women at higher risk (above 20% risk), once in a year. However, Women at moderate (below 15% & above 20% risk) or lower risk (below 15% risk) should consult their doctor before opting for Imaging tests [27].

2.3.2 Biopsy

Breast cancer can also be diagnosed by taking the sample [15] (biopsy) from the suspected tissue of the breast and examining it using the microscope. After the proper analysis of tissue sample, a pathologist can find both the presence of cancer as well as its type.

2.3.3 Clinical examination

It is nothing but the physical examination of the breast, armpits and collarbone by the doctor. It includes checking the swelling of the breast skin irritation, redness of breast skin, lump in the underarm etc. [29]

In our work, We are focusing only on the dataset obtained from the imaging tests of the breast. We are employing the different machine learning algorithms on the cell features obtained after the Imaging or screening tests.

Chapter 3

Hardware, Software and Dataset Used

3.1 Hardware Used

3.1.1 System Specifications

The full specifications of the system used in our work are given in Table 3.1.

TABLE 3.1: Table containing the datailed specification of Workstation

Workstation Model	HP Compaq Elite 8300
CPU	Intel ^{R} Core ^{TM} i7-3770
RAM	4 GB
Hard Drive	500 GB
Operating System	Ubuntu 18.04.02
Chipset	Intel ^{R} Q77 Express

3.1.2 CPU Specifications

The system is run by the Intel CoreTM i7-3770 processor. The detailed specifications of the CPU are shown in Table 3.2

TABLE 3.2: Table containing detailed specifications of CPU

CPU	Intel ^{R} Core ^{TM} i7-3770
# of Cores	4
# of Threads	8
Cache	8 MB SmartCache
Memory Type	DDR 1333
TDP	77 W
Processor Base Frequency	3.40 GHz

3.2 Software Used

Programming Language: Python Tool: Anaconda Navigator was used to implementing the Breast Cancer Diagnosis Model as it's a combination of the package manager and environment manager. It ensures all the dependencies for each package.

3.3 Libraries Used

Libraries: Numpy, Pandas, Scikit-Learn, Matplotlib

3.3.1 Numpy

Numerical Python, abbreviated as NumPy, developed in 2000, is a library of python for scientific computation. It is commonly used for multi-dimensional array processing. Along with it posses some other features such as: i) Broadcasting (sophisticated) functions ii) It can be used in Fourier transform, linear algebra etc. iii) It includes tools which can integrate C/C++ and Fortran code.

3.3.2 Pandas

Panda (derived from "panel data") is a python open source library released in 2008. It is employed for data analysis and manipulation. Specifically, it provides data structure and

operations to perform complicated tasks efficiently with a very small piece of code. It's built over the NumPy itself.

3.3.3 Scikit-Learn

It is an open source library built on the top of SciPy, Numpy and matplotlib. It includes various classification, clustering, regression and other algorithms which are useful in data analysis and mining.

3.3.4 Matplotlib

Matplotlib is the plotting library which is used to draw 2-D diagrams such as bar charts, histograms, plots, scatterplots, etc., with a small code.

3.4 Dataset

We are using Wisconsin Breast Cancer Data Set downloaded from the UCI repository [23]. It includes a total of 569 samples. Out of these 569 samples, 357 are benign rest 212 are malignant, as shown in Figure 1.

For the nucleus of each cell, 10 real-valued features [22] have been computed:

i) Radius - It is the mean of the distance between the centre of the cell and the points lying on its perimeter.
ii) Perimeter - It is the length of the outline/border of the breast cell.
iii) Texture - It is the standard deviation (SD) in grey-scale values. The formula for SD is given in Equation 3.1.

$$SD = \sqrt{(1/N)\sum_{j=1}^{N}(G_i - \mu)^2)} \tag{3.1}$$

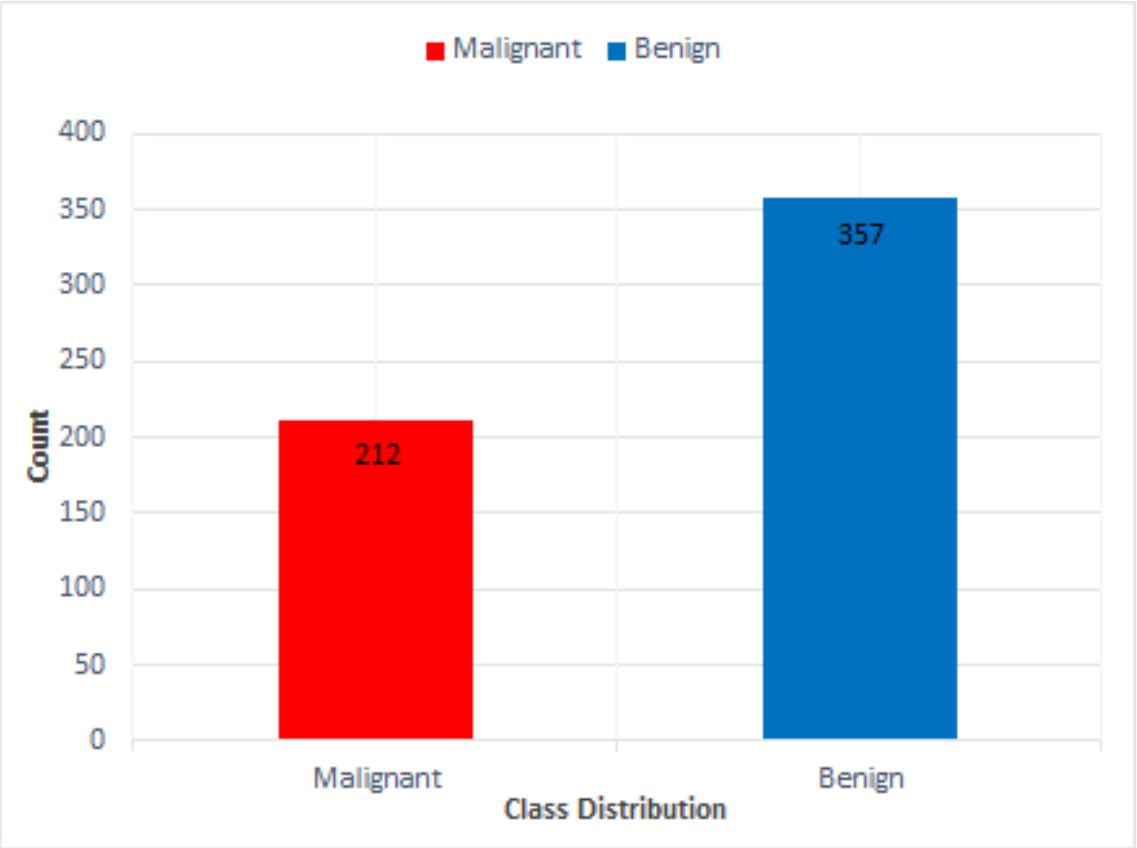

FIGURE 3.1: Wisconsin breast cancer dataset.

iv) Area - It is the average area of the breast cell.

v) Compactness - Cell compactness is calculated using Equation 4.1.

$$compactness = (\frac{perimeter^2}{area} - 1)g \tag{3.2}$$

vi) Concave points - It is the total number of concave portions of the contour of the cell.

vii) Smoothness - It is the local variation in the length of the radius of the cell, denoting how smooth the breast cell is.

viii) Symmetry

ix) Concavity - It shows the intensity of the concave portions of the cell contour, i.e. how concave the cell contour is.

x) Fractal dimension - It is the ratio providing a statistical index of complexity of pattern.

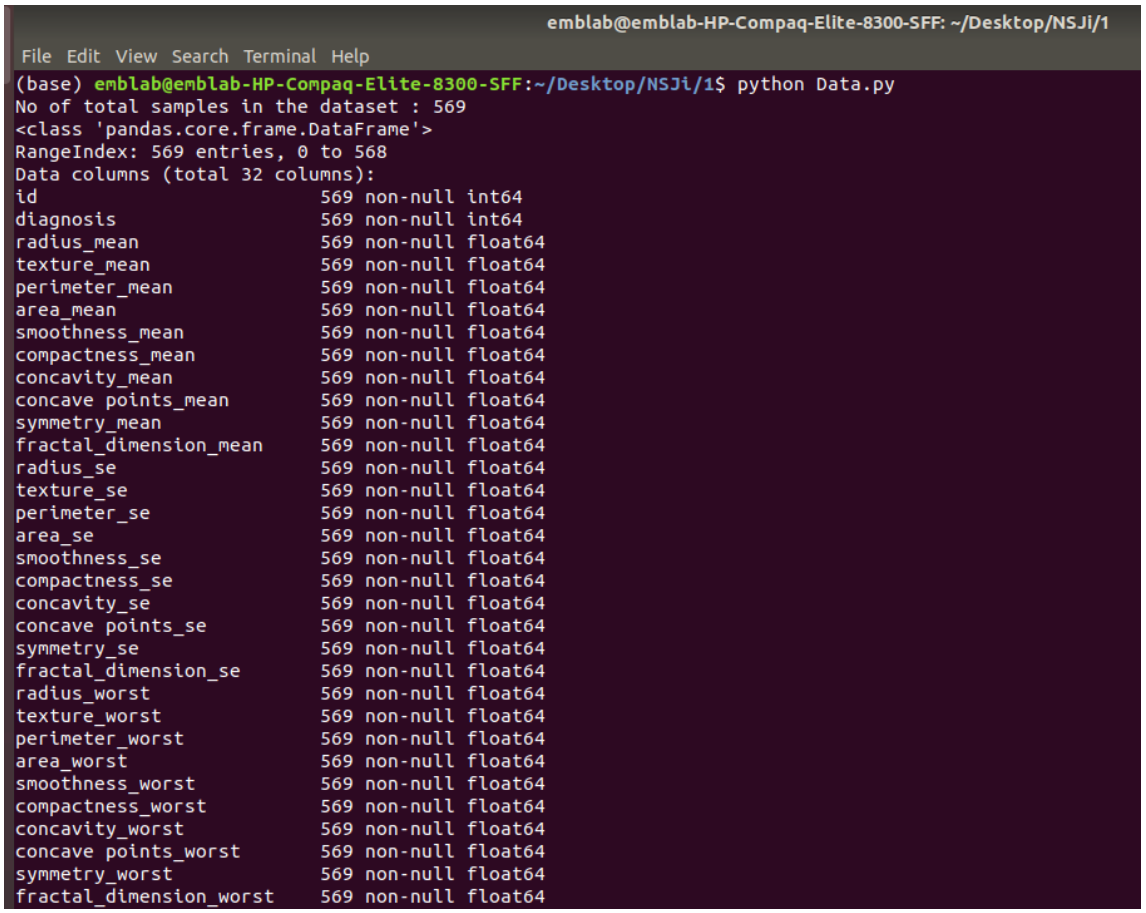

FIGURE 3.2: Description of the Dataset features

For all these 10 features of breast cell, the mean, standard error and worst values are calculated, resulting in overall 30 features, e.g. radius_mean, radius_se, radius_worst, texture_mean, texture_se, texture_worst, etc. All these features are computed to four significant digits after the decimal point in order to improve the accuracy in our work.

Chapter 4

Proposed Methodology

4.1 Supervised Machine Learning

Figure 4.1 shows the flowchart of the breast cancer diagnosis model using supervised machine learning. In supervised machine learning algorithms, a machine is trained using labelled data, such as an input where the desired output is known and based on this new data is classified. On processing the training data, the algorithm generates a mapping function which predicts the output for the new data after adequate training. In our work, the data is labelled as either malignant or benign. There are different methods such as regression, classification, gradient boosting and others which are used for predicting the output in supervised machine learning.

4.2 Machine Learning algorithms Used :

We have employed four machine learning algorithms, namely: Logistic Regression, KNN, SVM & naive Bayes. A brief description, along with their mathematical representation, is given below:

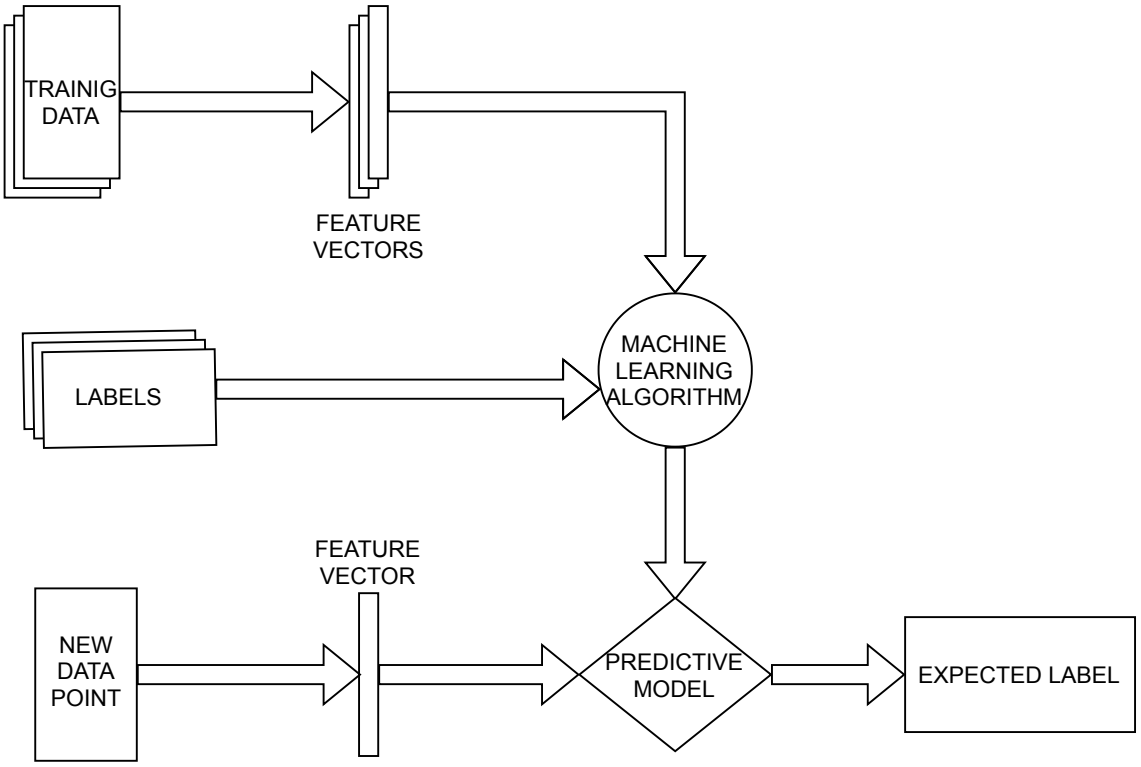

FIGURE 4.1: Flowchart of supervised machine learning.

4.2.1 Logistic Regression

Logistic regression [16] is a type of regression model. It predicts the dependent variable (result) which is categorical in nature, i.e. 0/1, pass /fail, yes/no etc., after finding a relation between the dependent variable and the given independent variables. Logistic regression uses the A sigmoid function given in Equation 4.1

$$y = \frac{1}{1 + e^{-}(x)} \tag{4.1}$$

for predicting the value of the dependent variable, which is dichotomous (binary) in nature. A sigmoid activation function is drawn in Figure 4.2.

The generalised equation for logistical regression is given in Equation 4.2.

$$y = \frac{1}{1 + e^{-}(b + c_1x_1 + c_2x_2 + c_3x_3 + ... + c_nx_n} \tag{4.2}$$

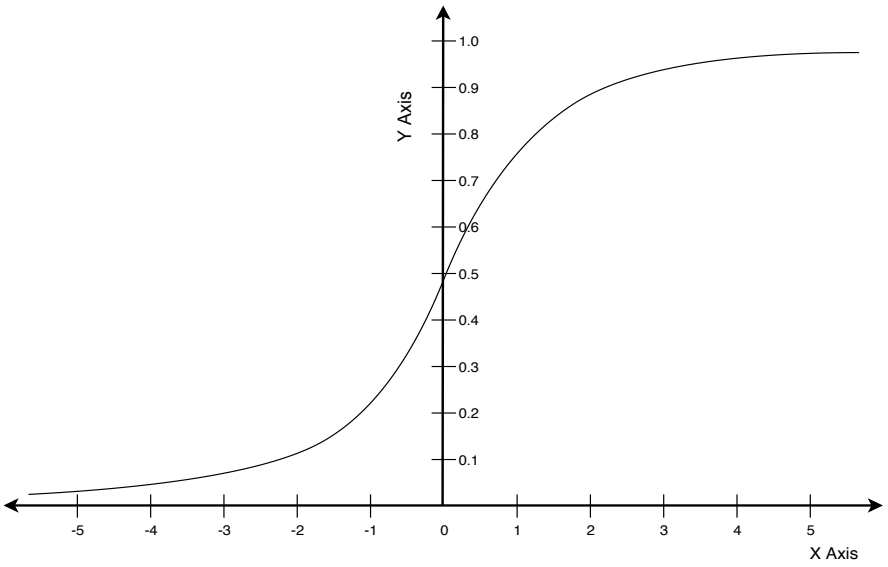

FIGURE 4.2: Sigmoid Activation Function.

where x_1,x_2,x_3, ... are independent variables, y is a dependent variable and b, c_1,c_2,c_3, ... are constants.

For our work, We have taken n=30 as there are 30 different cell features in the dataset.

4.2.2 Support Vector Machine

SVM [17], a supervised ML algorithm, is employed in both regression and classification. Mostly it's used in classification problems. In this ML algorithm, we plot each individual data as a point in the n-dimensional space where each dimension represents a particular feature of the dataset (also called the independent variable) as shown in Figure. A hyperplane or sets of hyperplanes are constructed for classifying different classes. The equation of hyperplane is $W^TX=0$, where W is the normal vector to the hyperplane.

Formally, a support vector machine finds a hyperplane or set of many hyperplanes in higher dimensional space, which can then be used for tasks like classification or outliers detection. A good separation is said to be achieved when the hyperplane found by the algorithm is such that it has the largest possible distance to the nearest training data-point of any of the class. This is the intuition behind the working of Support Vector Machines [7]. The working of the algorithm is beyond the scope of this work, hence intuition behind it is

only explained. Since the algorithm uses vectors (called as support vectors) to few data points of each class for finding the set of hyperplanes that divides those set of classes, it is known by the name of Support Vector Machines. The caveat here is that the SVM only tends to work when the data points are linearly separable. For the data points which are not linearly separable, the problem is looked in a much higher dimensional space which makes the separation of the classes easier, and SVM can then be applied. A plot of sample data points and their classification using SVM is shown in Figure 4.3. It can be seen from Figure 4.3 that H1 plane does a poor job in classifying the data points, H2 classifies the data points correctly but not in the most optimized way, and the H3 plane classifies the data points in the most optimal way. SVM have two disadvantages: choosing kenel function is cumbersome and longer training time as compared to other algorithms used.

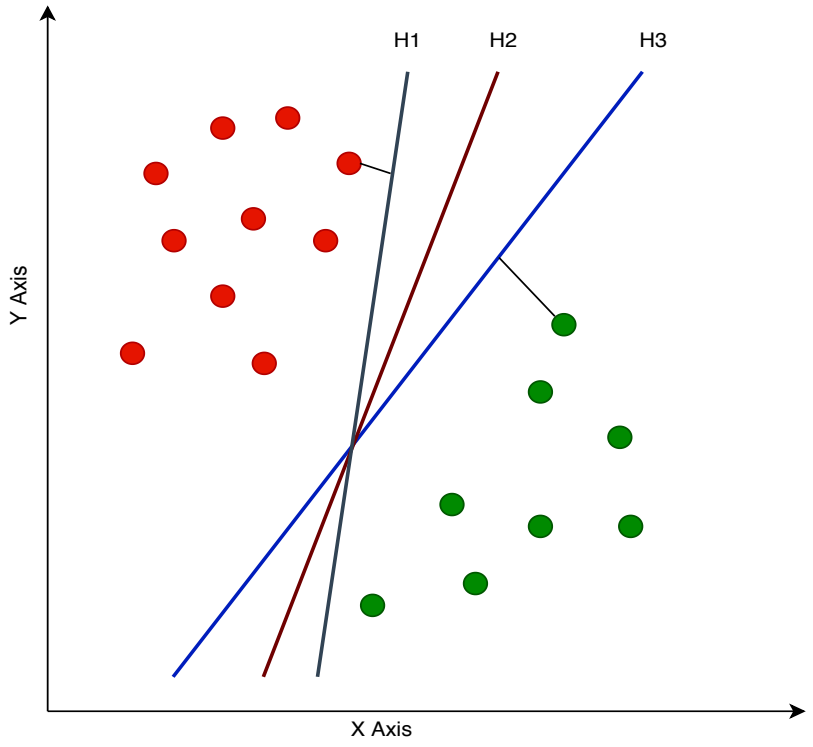

FIGURE 4.3: Support Vector Machine.

We have used radial basis function (RBF) kernel in our work. The mathematical representation of the RBF kernel [19] is given in Equation 4.3.

$$k(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{4.3}$$

4.2.3 k-Nearest Neighbours

k-nearest-neighbour [13] algorithm (KNN), is a type of supervised non-parametric machine learning algorithm. It classifies the new data point into one of the available categories based on the similarity principle or in another way it is the algorithm that assigns an unlabelled object, a label from a fixed set of labels. For example, a model can be trained based on the k-NN algorithm to classify the images into two classes square, circle. When any unclassified image is given, it will classify that image into either square or circle class.
The idea behind k-NN is straightforward. K number of nearest neighbours from the data point to be classified is taken from the set of objects for which the classification is already known. The class which will have the maximum number of neighbours points will be assigned as the label to the object under classification. k-NN [5] falls under the class of instance-based or lazy-learning algorithms, where the function is approximated locally, and all the computation for the classification phase is deferred. The k-NN algorithm is one of the most straightforward machine learning algorithms.

The class of the new data point is found out based on the majority vote [13] of its neighbouring data points. Number of the neighbours to be used for classification are decided manually. We have used the Euclidean distance our study to estimate the similarity. Euclidean distance [12] is calculated using the formula given in Equation 4.4.

$$EuclideanDisatance(x, x_i) = \sqrt{\sum x_j - x_{ij})^2} \tag{4.4}$$

Depending on the Euclidean distance, calculated with the help of above-given equation, k neighbours are selected and based on the majority vote of these k neighbours new data point is classified among one of the given classes. Figure 4.4 shows the k-nearest algorithms for two classes of datasets: class A and class B. Based on the value of k, we classify the new data point between one of them using majority voting.

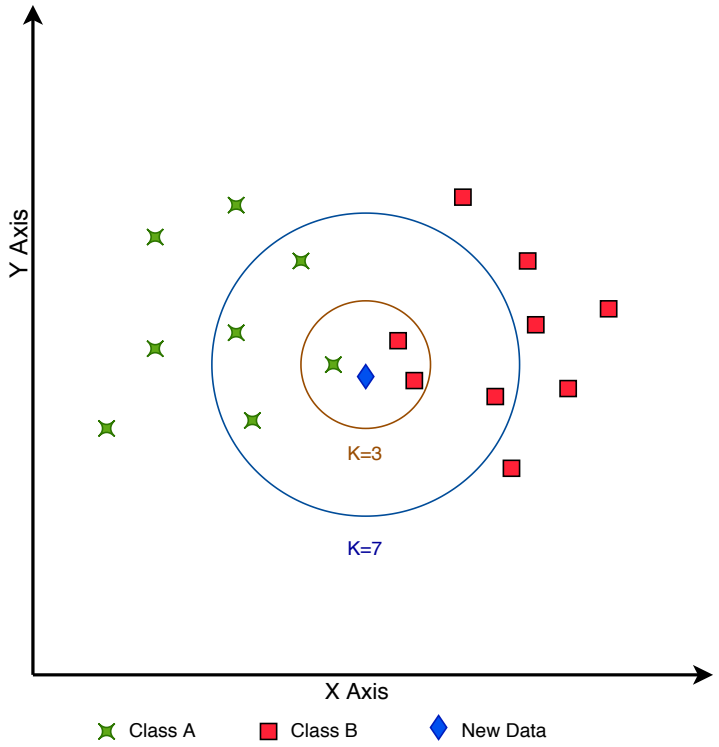

FIGURE 4.4: k-nearest neighbours.

4.2.4 Naive Bayes

A Naive Bayes classifier [10] is a probabilistic classifier model. It is used for binary (dichotomous) or multi-class classification problems. It is based on Bayes theorem. Bayes theorem is given in equation 4.5.

$$P(\frac{y}{X}) = \frac{P(\frac{X}{y})P(y)}{P(X)} \tag{4.5}$$

where y is dependent variable & X is a dependent feature vector of size n, shown in Equation 4.6.

$$X = (x_1, x_2, x_3, ..., x_n,) \tag{4.6}$$

Figure 4.5 shows a typical naive Bayes classifier which classifies the new data using conditional probability. Maximum a posteriori (MAP) decision rule has been used for constructing the classifier.

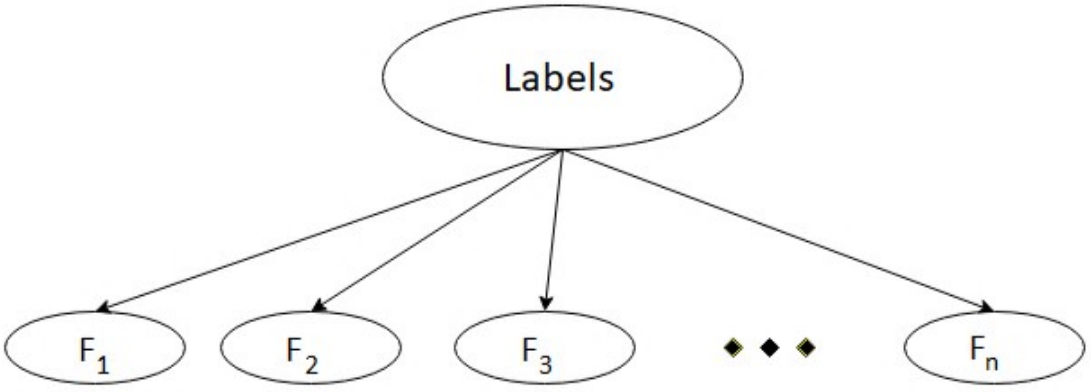

FIGURE 4.5: Naive bayes classifier.

Naive Bayes classifiers are of three types, namely, Gaussian naive Bayes Classifier, multinomial naive Bayes, Bernoulli naive Bayes. We will be using only Gaussian naive Bayes in this paper. Equation 4.7 shows the mathematical model of Gaussian function.

$$P(\frac{x_i}{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma^2{}_{\mathrm{y}}}) \tag{4.7}$$

4.3 Data Analysis

Wisconsin data set was partitioned into two subsets for training subset and testing subset, assigning different values for data selection in the algorithm manually. We estimated the parameters for all the four machine learning algorithms, namely: (1) Test Accuracy, (2) Sensitivity, (3) Specificity, (4) Positive predictive value (PPV) and (5) Negative predictive value (NPV).

Chapter 5

Model Training and Results

5.1 Model Training

For this work, the machine was trained on open-sourced models of all the four algorithms. We used Wisconsin dataset (.csv file) with object classes equal to 2, i.e. either malignant or benign and independent variables being equal to 30. We divided the dataset into two segments, one for training and another for testing. After the selection of training data, the model can be trained on them.

The model was trained on the HP Compaq Elite 8300 system with the Intel CoreTM i7-3770 processor. The detailed specification can be found in Tables 3.1 and 3.2. Each of the independent and dependent variables was fitted into the training model of all the four algorithms separately. After testing these models on the breast cancer dataset, accuracy and other parameters were found.

The diagram of the training model is given in Figure 5.1. The first step is about the data collection, which is to be diagnosed.

5.2 Results

We have constructed the proposed breast cancer diagnosis model using four different algorithms in python, namely: Logistic Regression, SVM, KNN Naive Bayes. We calculated

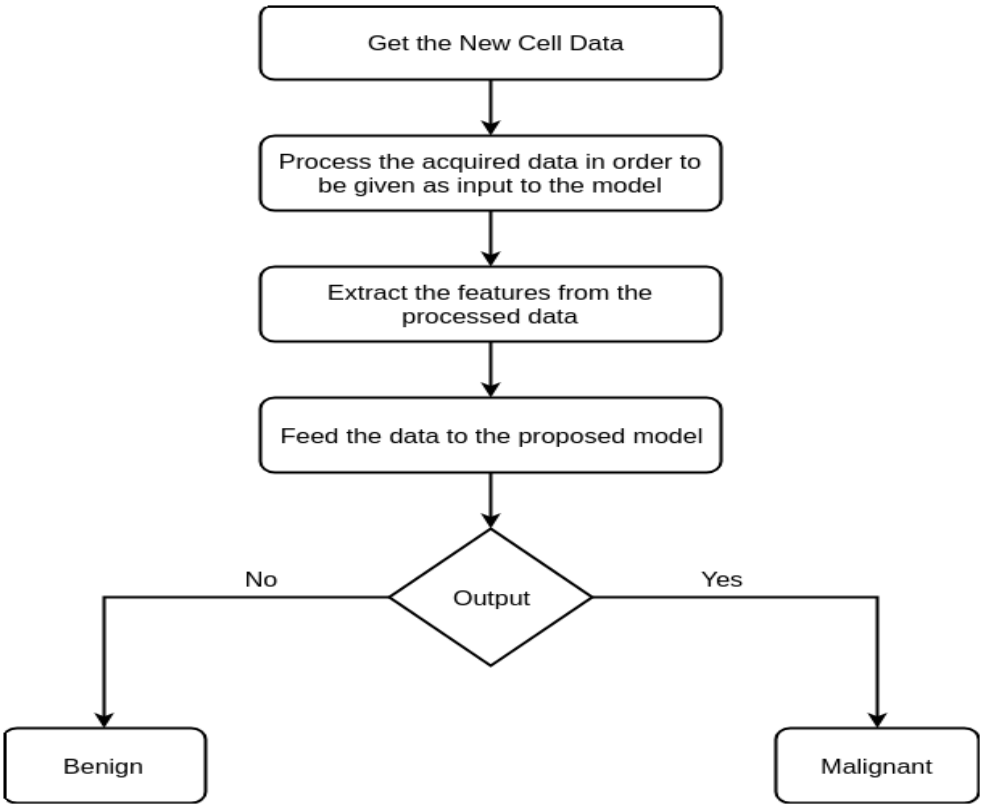

FIGURE 5.1: Model for breast cancer diagnosis.

the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each algorithm separately after varying their respective Hyper-parameters manually from the confusion matrix.

Suppose we have a confusion matrix as given in Table 5.1. Here W= Number of true positive, X= Number of false positive, Y=Number of false negative & Z=Number of true negative.

TABLE 5.1: Confusion Matrix

	Disease	Nondisease	Total
Positive	W(True Positive)	X(False Positive)	$T_{\text{Test Positive}}$
Neagative	Y(False Negative)	Z(True Negative)	$T_{\text{Test Negative}}$
	T_{Disease}	$T_{\text{Non-disease}}$	Total

A brief description of these parameters along with the formulas using which these are calculated(in terms of W, X, Y Z) is given below :

5.2.1 Accuracy

Accuracy is the measure of the correctness with which classifiers predict the exact class. Accuracy is calculated using Equation 5.1.

$$Accuracy = (\frac{W+Z}{W+X+Y+Z})X100\% \tag{5.1}$$

5.2.2 Sensitivity

Sensitivity is the probability with which the test predicts the presence of the disease in a person having the disease, calculated using Equation 5.2.

$$Sensitivity = \frac{W}{W+Y} \tag{5.2}$$

5.2.3 Specificity

Specificity is the probability with which the test predicts the absence of the disease in a person not having the disease, calculated using Equation 5.3.

$$Specificity = \frac{Z}{Z+Y} \tag{5.3}$$

5.2.4 PPV

Positive predictive value is the probability with which a person tests positive and that person is actually having the disease, calculated using Equation 5.4.

$$PPV = \frac{W}{W+X} \tag{5.4}$$

5.2.5 NPV

Negative predictive value is the probability with which a person tests positive and that person is actually not having the disease, calculated using Equation 5.5.

$$NPV = \frac{Z}{Z+Y} \tag{5.5}$$

All these parameters, i.e. sensitivty, specificity, PPV and NPV along with accuracy, were calculated from the confusion matrix and are given in Table 5.2.

TABLE 5.2: Comparison of accuracy, specificity, sensitivity, NPV and PPV for different Algorithms

Algorithms	***Accuarcy***	***Specificity***	***Sensitivity***	***PPV***	***NPV***
Logistic Regression	96.49 %	.9706	.9565	.9565	.9706
SVM	98.24 %	.9714	1.0	.9565	1.0
KNN	97.20 %	.9560	.9808	.9273	.9886
Naive Bayes	94.74 %	.9429	.9545	.9130	.9706

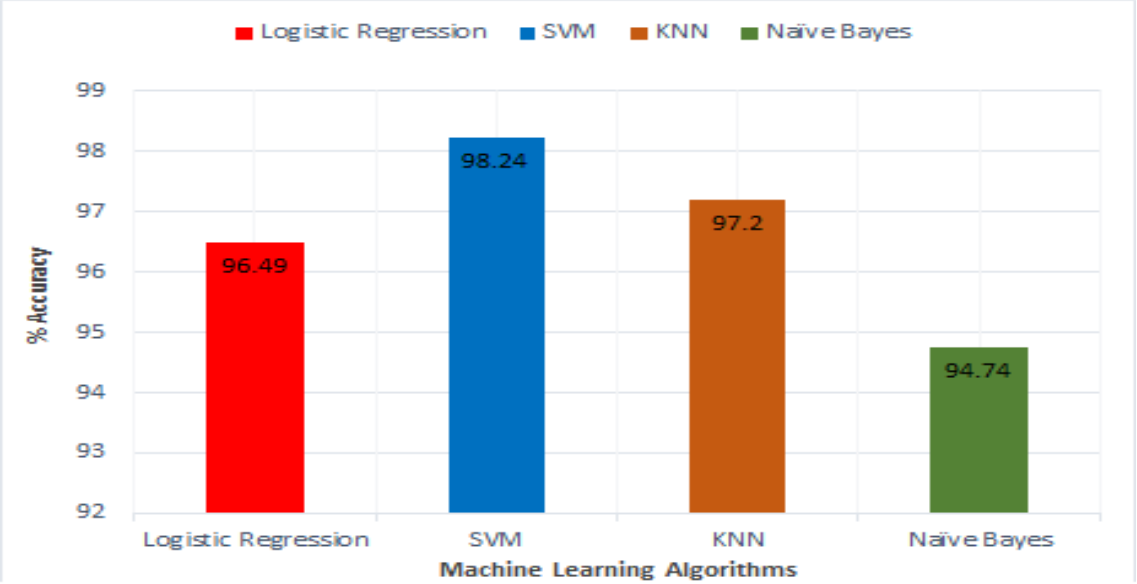

FIGURE 5.2: Accuracy for different Algorithms.

Fig 5.2 shows the bar chart comparing accuracies for all the four machine learning algorithms.

Chapter 6

Conclusion and Future Enhancements

6.1 Conclusion

In this work, the objective was to implement the breast cancer diagnosis model using machine learning. We used four different algorithms in our research. The model detected the breast cancer with satisfactory test accuracy and other parameters(specificity,sensitivity, NPV & PPV). After comparing these parameters of all the four algorithms, we can conclue that suport vector machine performed better than others. This model can help in early detection of breast cancer while reducing the dependency on the experts and minimizing the diagnosis cost & time.

6.2 Future Enhancements

This work diagnose breast cancer with good accuracy, but there are certain shortcomings which can be eradicated to improve the test accuracy. These required enhancements are given below :

- The system can be automated to send those data which are not classified correctly to the machine which will train the model further on these data for better results.
- The size of the dataset should be increased for improving the test accuracies and other parameters.
- Apart from these four algorithms, we can employ some other algorithms in future.

Bibliography

[1] Abien Fred M. Agarap. 2018. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18). ACM, New York, NY, USA, 5-9. DOI: https://doi.org/10.1145/3184066.3184080

[2] Shang Gao and Hongmei Li, "Breast cancer diagnosis based on support vector machine," 2012 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering, Jalarta, 2012, pp. 240-243.

[3] H. R. H. Al-Absi, B. Belhaouari Samir and S. Sulaiman, "A computer aided system for breast cancer detection and diagnosis," 2014 International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2014, pp. 1-4.

[4] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4.

[5] A. Mert, N. Kilic and A. Akan, "Breast cancer classification by using support vector machines with reduced dimension," Proceedings ELMAR-2011, Zadar, 2011, pp. 37-40.

[6] Z. Yan, H. Yanzhen and Y. Peng, "Computer Based Breast Cancer Diagnosis," 2009 Third International Symposium on Intelligent Information Technology Application Workshops, Nanchang, 2009, pp. 59-62.

[7] W. Yi and W. Fuyong, "Breast Cancer Diagnosis via Support Vector Machines," 2006 Chinese Control Conference, Harbin, 2006, pp. 1853-1856.

[8] H. Rajaguru and S. K. Prabhakar, "Expectation maximization based logistic regression for breast cancer classification," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 603-606. doi: 10.1109/ICECA.2017.8203608

[9] R. Radha and P. Rajendiran, "Using K-Means Clustering Technique to Study of Breast Cancer," 2014 World Congress on Computing and Communication Technologies, Trichirappalli, 2014, pp. 211-214.

[10] Cooper, 1999, An overview of the representation and discovery of causal relationships using Bayesian networks

[11] R. Delshi Howsalya Devi and P. Deepika, "Performance comparison of various clustering techniques for diagnosis of breast cancer," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, 2015, pp. 1-5.

[12] H. Jegou, M. Douze and C. Schmid, "Product Quantization for Nearest Neighbor Search," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, Jan. 2011. doi: 10.1109/TPAMI.2010.57

[13] D. O. Loftsgaarden C. P. Quesenbery "A nonparametric estimate of a multivariate density function" Ann. Math. Stat. vol. 36 pp. 1049-1051 June 1965.

[14] N. Alpaslan, A. Kara, B. Zencir and D. Hanbay, "Classification of breast masses in mammogram images using KNN," 2015 23nd Signal Processing and Communications Applications Conference (SIU), Malatya, 2015, pp. 1469-1472. doi: 10.1109/SIU.2015.7130121

[15] S. Singh, J. Harini and B. R. Surabhi, "A novel neural network based automated system for diagnosis of breast cancer from real time biopsy slides," International Conference on Circuits, Communication, Control and Computing, Bangalore, 2014, pp. 50-53.

[16] A. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," 2015 SAI Intelligent Systems Conference (IntelliSys), London, 2015, pp. 150-154. doi: 10.1109/IntelliSys.2015.7361138

[17] C. Cortes V. Vapnik "Support-vector networks" Machine learning vol. 20 no. 3 pp. 273-297 1995.

[18] B. Bektaş and S. Babur, "Machine learning based performance development for diagnosis of breast cancer," 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, 2016, pp. 1-4.

[19] X. Yang, H. Peng and M. Shi, "SVM with multiple kernels based on manifold learning for Breast Cancer diagnosis," 2013 IEEE International Conference on Information and Automation (ICIA), Yinchuan, 2013, pp. 396-399.

[20] https://timesofindia.indiatimes.com/india/indianowhas3rdhighestnumberofcancer-casesamongwomen/articleshow/60812041.cms

[21] https://nbcf.org.au/about-national-breast-cancer-foundation/

[22] https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin

[23] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

[24] https://en.wikipedia.org/wiki/Breast-cancer.

[25] https://radiology.ucsf.edu/sites/radiology.ucsf.edu/files/wysiwyg/research/RIGs/breast-cancer-rig/Breast-MRI.jpg

[26] http://www.radiologyassistant.nl/data/bin/

[27] https://www.cancer.org/cancer/

[28] https://www.medanta.org/

[29] https://www.breastcancer.org/sympotoms