

A Project Report

on

Using Supervised Learning for Breast Cancer Detection

Winter 2022-2023

Submitted in partial fulfillment of

the requirement for the award of the degree of

**Bachelor of Technology in Computer Science and
Engineering**



(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision

Dr. Kavita

Assistant Professor.

Submitted By

Group ID: BT4038

Name of Students

Jaideep Nagill - (19SCSE1010541)

Pratyaksh Singh - (19SCSE1010514)

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING /
DEPARTMENT OF COMPUTERAPPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
MAY, 2023**

Candidate's Declaration

We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **“Using Supervised Learning for Breast Cancer Detection”** in partial fulfillment of the requirements for the award of the School of Computer Science and Engineering submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, January 2023- June 2023, under the supervision of Name – Dr. Kavita assistant, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

Jaideep Nagill (19SCSE1010541)

Pratyaksh Singh (19SCSE1010514)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Kavita

Assistant Professor

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be impossible without the mention of the people who made it possible, whose constant guidance and encouragement crowned our efforts with success.

I would also like to thank Dr. Munish Sabharwal, Professor and Head, Department of Computer Science and Engineering, for his constant support.

I express my gratitude to Dr. Kavita, my project guide, for constantly monitoring the development of the project and setting up precise deadlines. Her valuable suggestions were the motivating factors in completing the work.

Finally, a note of thanks to the teaching and non-teaching staff of Dept of Computer Science and Engineering, for their cooperation extended to me, and my friends, who helped me directly or indirectly in the course of the project work.

CERTIFICATE

The Final Project Viva-Voce examination of Jaideep Nagill– 19SCSE1010541 & Pratyaksh Singh– 19SCSE1010514 has been held on _____ and their work is recommended for the award of Bachelor of Technology.

Signature of Examiner(s)

Signature of Supervisor(s)

Signature of Program Chair

Signature of Dean

Date: May, 2023

Place: Greater Noida

Abstract

The most commonly causing cancer among Indian women is breast cancer and it affecting all over world with its impact. According to the medical reports of breast cancer patients in India were unable to hold the pain and about half of them are dying. In the proposed work used a machine learning algorithm to decrease the pre-processing time and to detect the symptoms and for better accuracy.

The system is trained pre-processed image of fed to the system which are in the form of mammograms in common the X-ray of breast. The system which has the data segregated into the training and testing datasets analyses the input images based on the characters or the labels assigned to them done with the application of few of the algorithms which are present in the machine learning we compare the data or the image and probable output based on the character labels is obtained in the form of result. Compared to existing work and the proposed machine learning model as a series of combination and permutations of algorithms lead to increase in the efficacy of the result and got the accuracy of 97.4% using random forests algorithm.

Contents

S. No.	Particulars	Page No.
1	Introduction	8
2	Literature Survey	11
3	Visualizing the Data	19
4	Model Selection	20
5	Result	24
6	Future Scope	25
7	Conclusion	26
8	References	27

List of Figures

Figure 1: Data Flow Diagram for breast cancer diagnosis.....	16
Figure 2: Flow Chart flow Diagram Medical Imaging Breast imaging.....	17
Figure 3: Use case Diagram.....	18
Figure 4: Heat Map	19
Figure 5: Confusion Matrix	23

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Introduction

Breast cancer signifies unique of the diseases that as more losses each year. Breast cancer is the utmost collective cancer amongst women universal secretarial for 25% of all cancer cases and pretentious 2.1 million persons in 2015 primary diagnosis suggestively rises the likelihoods of persistence. The existing methods are Machine learning, method of training machines with data to make the decision for same conditions and its application can be observed in various domains such as medical, network, object identification and security etc. There are 2 machine learning types that is single and hybrid approaches as for instance Support vector machine. Breast Cancer is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. In the year 2018 according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women”

The main tribute of using machine learning in early breast cancer detection is to enable the prediction and improving accuracy of decision making. By using this machine learning the tumor can be identified as malignant or benign hence the unnecessary surgeries and painful operations can be decreased. The machine learning can be more accurate by providing more dataset. Machine learning doesnot require human intervention it gives the ability for the machine to learn on itsown. These algorithms increase accuracy and effectiveness as the machine gains experience this helps to get the better decision outcome. Breast cancer detection using machine learning has achieved successfully with accuracy up to 97.4%. By using this machine learning the output is effective and faster and reduces the complexity. Here we have used combination of classifiers & algorithms such as decision tree algorithm, random algorithm and logistic regression helped to achieve high accurate and efficient model some of the image compression techniques explained based on region on interest. In 2020 30% of newly diagnosed cancer in women as per the survey. Proposed work used the

supervised machine learning algorithm used for detection. Here three algorithms used namely Logistic Regression, Decision tree algorithm, Random forests algorithm.

This research paper has gathered information from ten different papers based on breast cancer using machine learning and other techniques such as ultrasonography, blood analysis etc. The project by S. Gokhale, is using the ultrasonography (USG) technique which is a powerful method in detecting details about the breast mass that usually cannot be detected even by mammography.

Another project by Pragya Chauhan and Amit Swami, which is based on the ensemble method usually used to increase the prediction accuracy of breast cancer. A Genetic algorithm based weighted average method that includes crossover and mutation is used for the prediction of multiple models.

Furthermore, a project by Abien Fred M. Agarap uses different methods like GRU-SVM, NN, multilayer perceptron (MLP), SoftMax regression to classify the dataset into benign or malignant. A project by Priyanka Gupta shows the comparison of the lesser invasive techniques such as Classification and Regression Trees (CART), random forest, nearest neighbor and boosted trees. These four classification models are chosen to extract the most accurate model for predicting cancer survivability rate. It draws results that are from methods like Extreme Learning Machine (ELM), ANN etc. it also has an added MATLAB GUI environment that for classification with ANN. Furthermore, a performance evaluation using three indicators including prediction accuracy values, F-measure metric and AUC values are used to compare the performance of these five classification models. Other experiments show that random forest model can achieve better performance and adaptation than other four methods. A project by Mumine Kaya Keles, which is a comparative study of data mining classification algorithms. Another project by Sang Won Yoon and Haifeng Wang that uses four data mining models are applied in this paper, i.e., support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree. Furthermore, feature space is highly deliberated in this paper due to its high impact on the

efficiency and effectiveness of the learning process.

Lastly a project by Wenbin Yue and Zidong Wang that shows the algorithms that helped them with the diagnosis and prognosis of their dataset.

Literature Survey

Ultrasound characterization of breast masses by S. Gokhale written by proposed asystem where they found that doctors have known and experienced that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, forming a lump or mass that may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body.

Some women can be at a higher risk for breast cancer because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment. In this study we use four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbor, Support Vector Machine, Artificial Neural Network and random forest. Harmonic imaging and real-time compounding have been shown to enhance image resolution and lesion characterization. More recently, USG elastography seems to be quite encouraging. Initial results show that it can improve the specificity and positive predictive value of USG within the characterization of breast masses. The reason why any lesion is visible on mammography or USG is that the relative difference within the density and acoustic resistance of the lesion, respectively, as compared to the encompassing breast tissue.

[1] Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach written by P. Chauhan and A. Swami proposed a system where they found that Breast cancer prediction is an open area of research. In this paper different machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest, support vector machine, neural network, linear model, adaboost, naive bayes methods are used for prediction. An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which overcame the limitations of the classical weighted average

method. Genetic algorithm based weighted average method is used for prediction of multiple models. The comparison between Particle swarm optimization (PSO), Differential evolution (DE) and Genetic algorithm (GA) and it is concluded that the genetic algorithm outperforms for weighted average methods. One more comparison between classical ensemble method and GA based weighted average method and it is concluded that GA based weighted average method outperforms.[2] On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset by the A. Fred and M. Agarap. In this paper, six machine learning algorithms are used for detection of cancer. GRUSVM model is used for the diagnosis of breast cancer GRUSVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search, SoftMax Regression, and Support Vector Machine (SVM) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by measuring their classification test accuracy, and their sensitivity and specificity values. The said dataset consists of features which were computed from digitized images of FNA tests on a breast mass. For the implementation of the ML algorithms, the dataset was partitioned in the following fashion 70 percent for training phase, and 30 percent for the testing phase. Their results were that all presented ML algorithms exhibited high performance on the binary classification of carcinoma, i.e. determining whether benign tumors or malignant tumor. Therefore, the statistical measures on the classification problem were also satisfactory. To further corroborate the results of this study, a CV technique such as k-fold cross-validation should be used. The appliance of such a way won't only provide a more accurate measure of model prediction performance, but it'll also assist in determining the foremost optimal hyper-parameters for the ML algorithms.[3] Analysis of Machine Learning Techniques for Breast Cancer Prediction by the P. Gandhi and Prof. Shalini L of VIT university, Vellore. In this paper, ML techniques are explored in order to boost the accuracy of diagnosis. Methods such as CART, Random Forest, K-Nearest Neighbors are compared. The dataset used is acquired from UC Irvine Machine Learning Repository. It is found that KNN algorithm has much better performance than the

other techniques used in comparison. The most accurate model was K- Nearest Neighbor. The classification model such as Random Forest and Boosted Trees showed the similar accuracy. Therefore, the most accurate classifier can be used to detect the tumor so that the cure can be found in early stage.[4] Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data by the M. Fatih Aslan, Y. Celik, K. Sabanci, and A. Durdu for carcinoma early diagnosis. During this paper, four different machine learning algorithms are used for the early detection of carcinoma. The aim of this project is to process the results of routine blood analysis with different ML methods.

Methods used are Artificial Neural Network (ANN), Extreme Learning Machine (ELM), Support Vector Machine (SVM) and Nearest Neighbor (k-NN). Dataset is taken from the UCI library. In this dataset age, BMI, glucose, insulin, homeostasis model assessment (HOMA), leptin, adiponectin, resistin, and chemokine monocyte chemoattractant protein (MCP1) attributes were used.

Parameters that have the best accuracy values were found by using four different Machine Learning techniques. This dataset includes age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1 features that can be acquired in routine blood analysis. The significance of these data in breast cancer detection was investigated by ML methods. The analysis was performed with four different ML methods. k-NN and SVM methods are determined using Hyperparameter optimization technique. The highest accuracy and lowest training time were given by ELM which was 80% and 0.42 seconds.[5] Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction by Yixuan Li and Z. Chen used two datasets in the study. The study firstly collects the data of the BCCD dataset which contains 116 volunteers with 9 attributes and data of WBCD dataset which contains 699 volunteers and 11 attributes. Then we preprocess the raw data of WBCD dataset and obtained the info that contains 683 volunteers with nine attributes and therefore the index indicating whether the volunteer has the malignant tumor. After comparing the accuracy, F Measure metric and ROC curve of 5 classification models, the result has shown that RF is

chosen as the primary classification model during this study. Therefore, the results of this study provide a reference for experts to distinguish the character of carcinoma. In this study, there are still some limitations that ought to be solved in further work. For instance, though there also exist some indices people haven't found yet, this study only collects the info of 10 attributes during this experiment.[6] Artificial Neural Networks in Image Processing for Earlier Detection of Breast Cancer by M.M.Mehdy, E.E.Shair and P.Y.Ng, Computational and Mathematical Methods in Medicine, (SVM), Artificial neural network m (ANN), Gaussian mixture model (GMM), Linear regressive classification (LRC), K- Nearest neighbor (KNN), Weighted hierarchical adaptive voting ensemble (WHAVE), etc.

Classification and data mining method is an actual way to categorize statistics.[7] Mammogram of Breast Cancer Detection Based Using Image Enhancement Algorithm by V. kumar , K.Patel, Prof.S. Uvaid and Prof.A.C.Suthar Categorizing data with reverence to competence and proficiency of individually algorithm in relation to precision, accurateness, intuition and specificity is done. Outcomes illustration that SVM gives the utmost accuracy with lowest error rate.[8] Solving Mammography Problems of Breast Cancer Detection Using Artificial Neural Networks and Image Processing Techniques by Moh'd Rasoul A Al-Hadidi, Mohammed Y. Al- Gawagzeh the proposed methods. Here, 569 patients' images were used for analysis, who had a wide range of 33 class labels assigned to them. Each class label is a parameter of the cancerous non- cancerous cells, and we make an attempt to predict the cancer just by the input image of scanned breast image. The approach described states the image is resized to required number of pixels, and we perform both the model optimization and predictions on these down-scaled images.[9] Cole, E. B. et al. Impact of Computer-Aided Detection Systems on Radiologist Accuracy With Digital Mammography. by Am. J. Roentgenol these are broadly used in identification and analysis to sort results. Support Vector Machine, Decision Tree, Naive Bayes and k-NN on the Wisconsin Breast Cancer (original) datasets is accompanied.[10] Fenton, J. J. et al. by Influence of Computer-Aided Detection on Performance of Screening Mammography The machine learning can be

more accurate by providing more dataset. Machine learning does not require human intervention it gives the ability for the machine to learn on its own. These algorithms increases accuracy and effectiveness as the machine gains experience this helps to get the better decision outcome. Breast cancer detection using machine learning has achieved successfully with accuracy up to 97.4%. By using this machine learning the output is effective and faster and reduces the complexity. Here we have used combination of classifiers & algorithms such as decision tree algorithm, random algorithm and logistic regression helped to achieve high accurate and efficient model.[11]histopathologyImage Analysis Using Image Processing Technique, Signal & Image Processing byA. D. Belsare and M. M. MushrifHThis model uses the clinical dataset which contains images of mammogram of around 569 patients the dataset which is used for the preprocessing has been assimilated in the usage of mammograms and thenfor the Saturn parameters which are stated below to acquire these parameters the pre-processing has been implemented then obtain data set has been divided into training data set under supervise machine learning models this medical image dataset has been then distributed into practice and testing data sets for the purposeof evaluation and validation of the project the training data group is further cut into the feature vectors weather in to the machine learning algorithm a predictive.

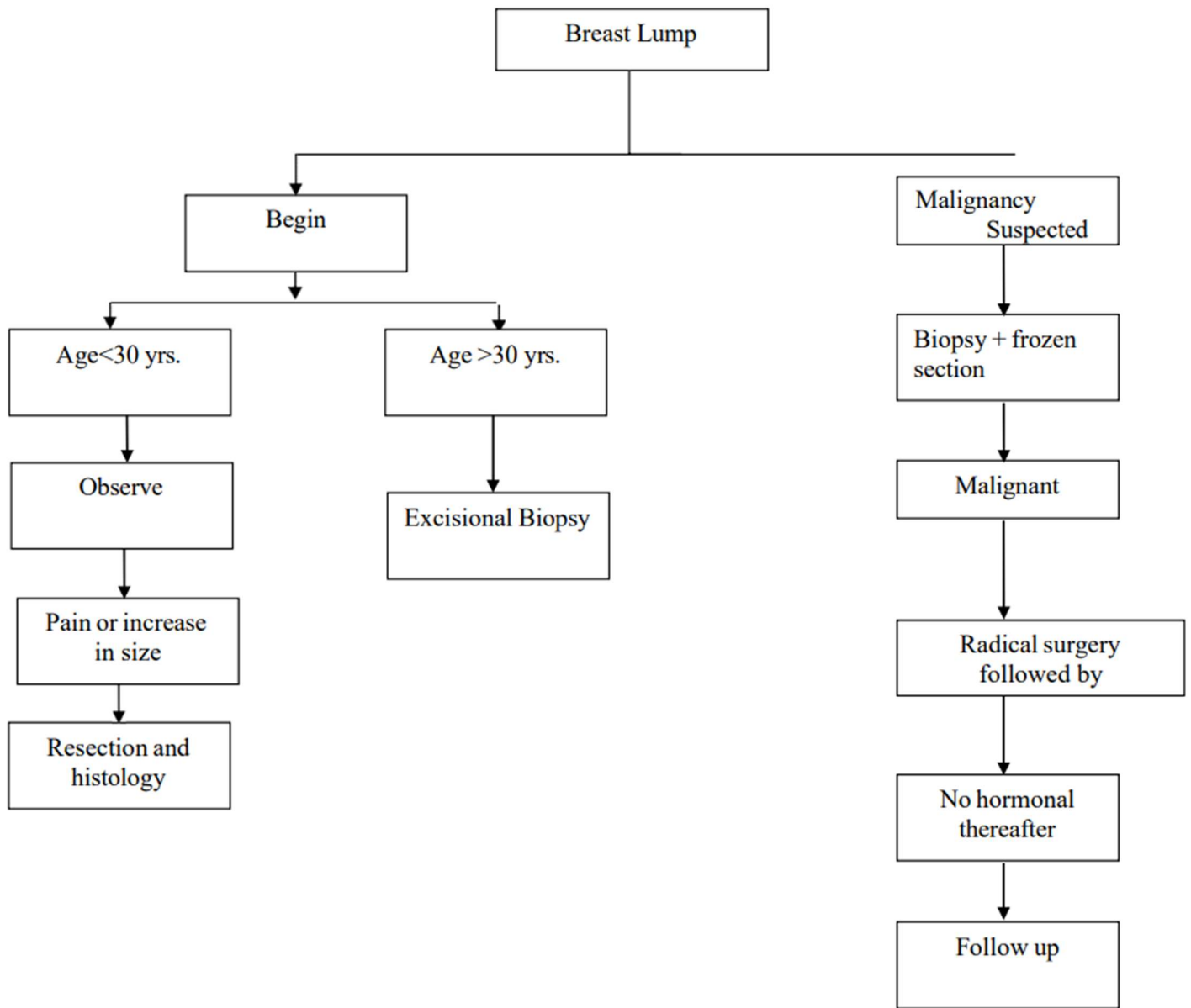


Figure 1: Data Flow Diagram for breast cancer diagnosis

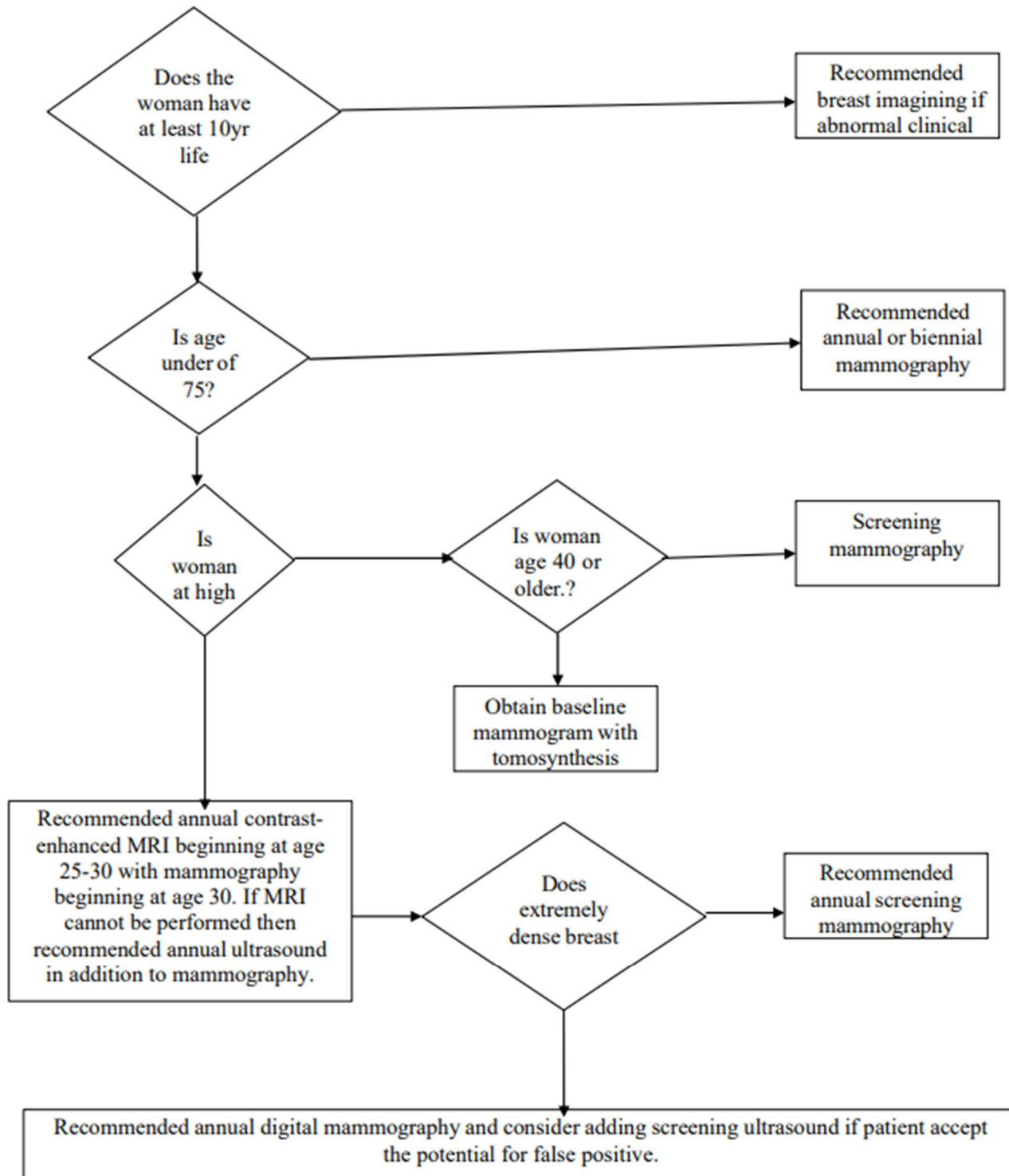


Figure 2: Flow Chart flow Diagram Medical Imaging Breast imaging

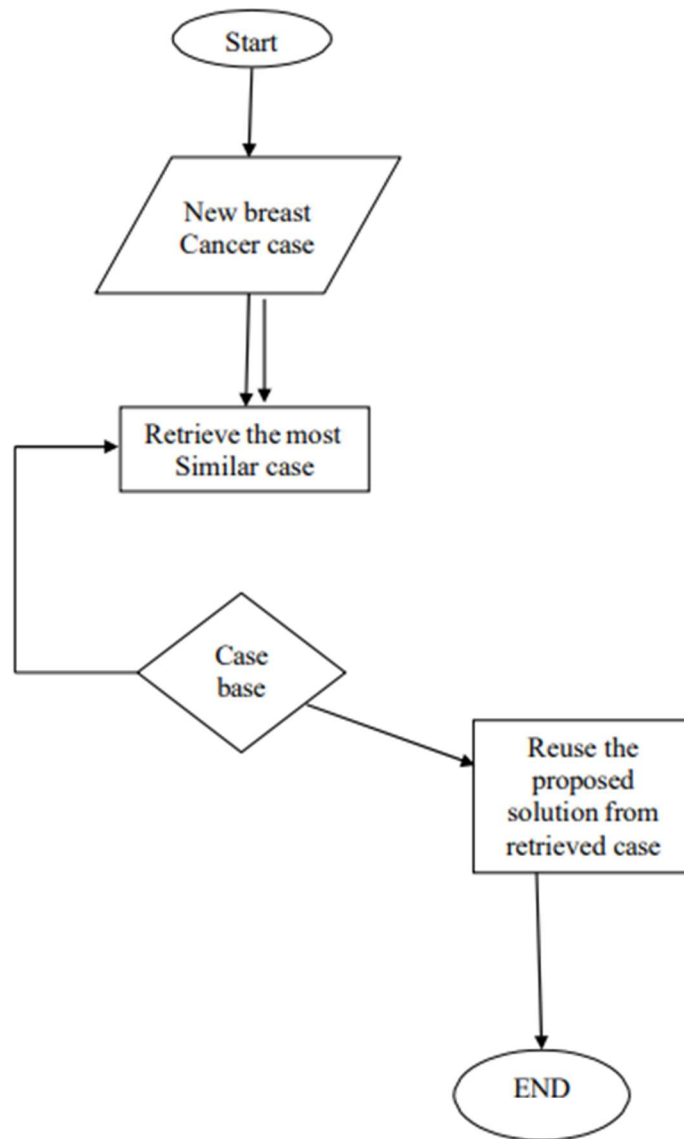


Figure 3: Use case Diagram

Visualizing the Data

We need to build visualizations of the data in order to decide how to proceed with the machine learning tools. The [Seaborn](#) and the [Matplotlib](#) packages will be used for this purpose. We use the mean values of the features. So first we will have to separate those features in the list to make some work easier and the code more readable

```
features_mean= list(data.columns[1:11])
```

The first method that can be used for visualization is heat map. A heat map is a two-dimensional representation of data in which values are represented by colors. A simple heat map provides an immediate visual summary of information. More elaborate heat maps allow the viewer to understand complex datasets.

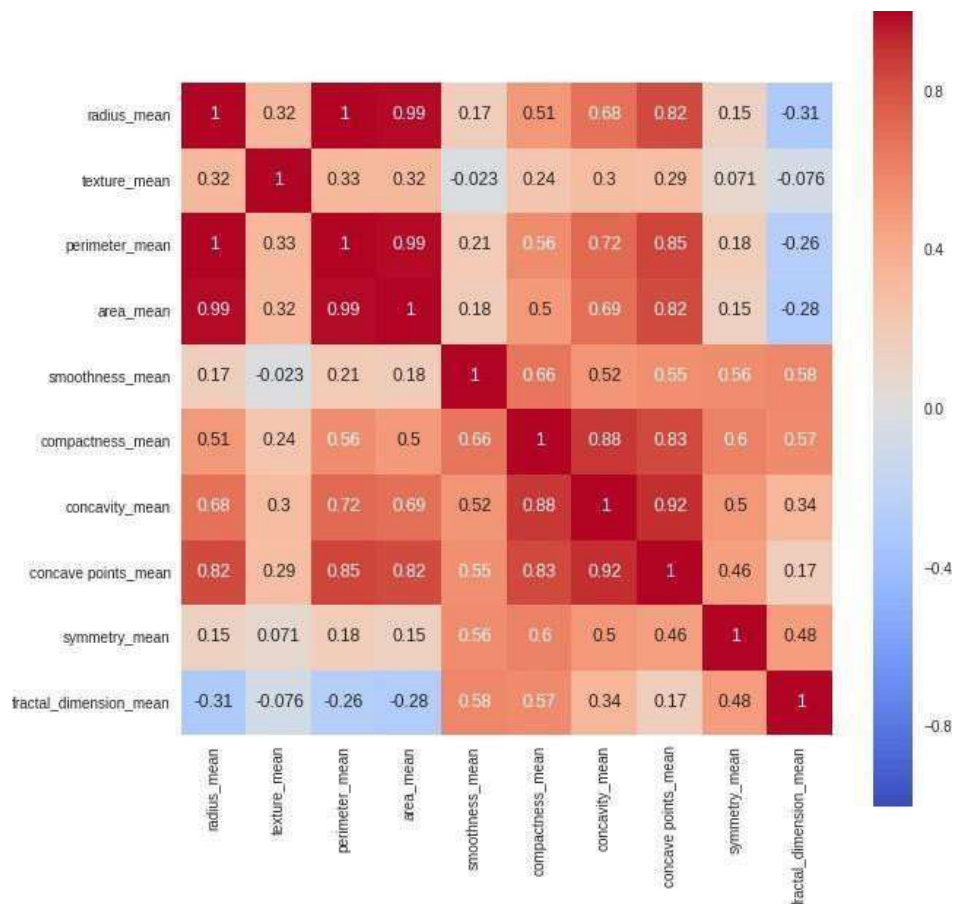


Figure 4: Heat Map

Model Selection

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results.

Usually, Data Scientists use different kinds of Machine Learning algorithms to the large data sets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning.

Without wasting much time, I would just give a brief overview about these two types of learnings.

Supervised learning: Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into **Regression** and **Classification** problems.

A **regression** problem is when the output variable is a real or continuous value, such as “salary” or “weight”.

A **classification** problem is when the output variable is a category like filtering emails “spam” or “not spam”.

Unsupervised Learning: Unsupervised learning is the algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B(Benign). So, we will use Classification algorithm of supervised learning.

We have different types of classification algorithms in Machine Learning: -

1. Logistic Regression
2. Nearest Neighbor
3. Support Vector Machines
4. Kernel SVM
5. Naïve Bayes
6. Decision Tree Algorithm
7. Random Forest Classification

Lets start applying the algorithms:

We will use sklearn library to import all the methods of classification algorithms.

We will use Logistic Regression method of model selection to use Logistic Regression Algorithm,

```
#Using Logistic Regression Algorithm to the Training Set from sklearn.linear_model
import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, Y_train)#Using KNeighborsClassifier Method of neighbors
class to use Nearest Neighbor algorithm from sklearn.neighbors import
KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
```

```

classifier.fit(X_train, Y_train)
#Using SVC method of svm class to use Support Vector Machine Algorithm
from sklearn.svm import SVC
classifier = SVC(kernel = 'linear', random_state = 0)
classifier.fit(X_train, Y_train)
#Using SVC method of svm class to use Kernel SVM Algorithm
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, Y_train)
#Using GaussianNB method of naive_bayes class to use Naïve Bayes Algorithm
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, Y_train)
#Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, Y_train)
#Using RandomForestClassifier method of ensemble class to use Random Forest
Classification algorithm
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy',
random_state = 0)
classifier.fit(X_train, Y_train)

```

We will now predict the test set results and check the accuracy with each of our model:

```
Y_pred = classifier.predict(X_test)
```

To check the accuracy we need to import confusion matrix method of metrics class. The confusion matrix is a way of tabulating the number of mis- classifications, i.e., the number of predicted classes which ended up in a wrong classification bin based on the true classes.

```

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred)

```

We will use Classification Accuracy method to find the accuracy of our models. Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$\text{Accuracy} = \frac{\text{Number of Correct predictions}}{\text{Total number of predictions made}}$$

To check the correct prediction, we have to check confusion matrix object and add

the predicted results diagonally which will be number of correct prediction and then divide by total number of predictions.

	0	1
0	87	3
1	3	50

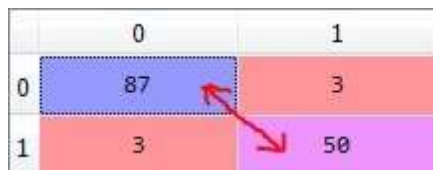


Figure 5: Confusion Matrix

After applying the different classification models, we have got below accuracies with different models:

1. Logistic Regression — 95.8%
2. Nearest Neighbor — 95.1%
3. Support Vector Machines — 97.2%
4. Kernel SVM — 96.5%
5. Naive Bayes — 91.6%
6. Decision Tree Algorithm — 95.8%
7. Random Forest Classification — 98.6%

So finally, we have built our classification model and we can see that Random Forest Classification algorithm gives the best results for our dataset. Well its not always applicable to every dataset. To choose our model we always need to analyze our dataset and then apply our machine learning model.

Result

In this paper, detection of breast cancer was done using DNN algorithm. There are several other algorithms are also available such as ANN, SVM, KNN etc. But as per the accuracy level is concerned, DNN provides a better result, using DNN algorithm accuracy and level of sensitivity is increased than another algorithm.

Thus, the earlier detection of breast cancer is detected without any physical contact. This method avoids destruction of the part being tested.

This paper uses the wavelet transformation and GLCM feature extraction. These features were useful to distinguish the maximum number of samples accurately. Finally, the simulated results show that used methodologies provides better classification rate with minimum error rate for all test samples.

Future Scope

The Future of Cancer Prognosis Look Like: -

AI is set to change the medical industry in the coming decades — it wouldn't make sense for pathology to not be disrupted too.

Currently, ML models are still in the testing and experimentation phase for cancer prognoses. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models.

Here's what a future cancer biopsy might look like:

You perform clinical tests, either at a clinic or at home. Data is inputted into a pathological ML system. A few minutes later, you receive an email with a detailed report that has an accurate prediction about the development of your cancer.

While you might not see AI doing the job of a pathologist today, you can expect ML to replace your local pathologist in the coming decades, and it's pretty exciting.

Conclusion

In this paper we have worked to collect the suitable dataset needed to help in this predictive analysis. This dataset is then processed to remove all the junk data. The predictive analysis method is being used in many different fields and is slowly picking up pace. It is helping us by using smarter ways to solve or predict a problem's outcome. Our scheme was developed to reduce the time and cost factors of the patients as well as to minimize the work of a doctor. We have tried to use a very simple and understandable model to do this job. Next, machine learning algorithms should be used on the training data and the testing data should be used to check if the outcomes are accurate enough.

References

- [1] Wolberg, Street and Mangasarian, "Wisconsin Diagnostic Breast Cancer Dataset"
<http://archive.ics.uci.edu/ml>
- [2] Mengjie Yu, "Breast Cancer Prediction Using Machine Learning Algorithm", The University of Texas at Austin, 2017.
- [3] Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne and Xiaohui Liu,
- [4] "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", 2018.
- [5] S. Palaniappan and T. Pushparaj, "A Novel Prediction on Breast Cancer from the Basis of Association rules and Neural Network", 2013.
- [6] Joseph A. Cruz, David S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis"