Breast Cancer Relapse Prognosis using Machine Learning Approach

Dr. Kamalakshi Naganna[1] Pooja A[2] Sushma D[3] Rashmi R V[4] Rachana R[5]
[1]Professor & Head [2,3,4,5]UG Student
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Sapthagiri College of Engineering, Bengaluru, India

Abstract— Cancers diseases are massive trouble in the community. Breast cancer is dominant reason for death amidst women. It presents the study of various machine learning approaches to predict breast cancer recurrence. The prime intention is to focus on advantages and disadvantages of classification techniques. It compares the results achieved by each machine learning algorithms and predicts recurrence of breast cancer.

Keywords: Breast Cancer, Classifiers, Recurrence, Relapse, Logistic, Deep Neural Network, SVM, K-Means

I. INTRODUCTION

Breast Cancer stands as second most cancer found in women after skin cancer. The tumour prediction is study that deals with three areas which are susceptibility, recurrence, and survivability. Mammogram technique can be used to detect cancer as early as possible. Since the breast cancer is hereditary which are possibly to inherit through genes, there has been genes examination in order to know beforehand.

Comparatively the survival rate of women with breast cancer in India is 66.1% which is less when compared to the women in America and Australia where the survival rate is high with 90%. The prime reason for inferior survival rates of breast cancer in India is that people in India are less aware of the information of the breast cancer and its treatment. This treatment is troublesome if the people visit doctors at final stage. The method or technique used in diagnosing in India is not efficient as in other countries.

Women who had diagnosed with breast cancer will have the fear of reappearance of cancer. When Breast cancer comes again, it's called recurrence. Usually breast cancer will re-appear within first five years after treatment. Breast cancer will come again either as local recurrence or in other parts of the body. Re appearance of breast cancer might occur in bones, lymph node outside the breast. Patient data or information, it is possible to predict if breast cancer will come again or not. Doctors can use different kinds of techniques in-order to treat patients if they know whether breast cancer will come again or not. Machine learning algorithms can be used to predict the re-appearance of breast cancer. These algorithms can be implemented in genetic as well as clinical data. Usage of machine learning algorithms can increase the accuracy of prediction.

II. LITERATURE SURVEY

Wei Wang and et.al [1] explained the usage of kernel principal component analysis for feature extraction and the support vector machine to classify the statuses of through-wall human being detection. Here the importance of the KPCA is emphasized which can even perform detection of more than one status and also it will be able to perform pattern recognition of non-linear data. Principal component analysis (PCA) in general can be stated as method of linear dimensionality reduction used as feature extraction for large amount of data. PCA performs effectively based on certain observations and these differ precisely and will be defined by using second-order correlations. Based on kernel function KPCA can be concluded as non-linear PCA. The accuracy of the KPCA is 81.75% and is significantly higher than PCA algorithm which is 79.125%. KPCA has kappa value more than 0.75 with better consistency, and PCA has kappa value less than 0.75 consistency. Mean, root mean square and absolute mean error is calculated using anticipated result of the classification algorithm. In this way, performance is determined, it is found that KPCA is efficient than PCA.

Sandeep Kaur and et.al [2] provided a survey on various feature extraction methods like Genetic, SVM-RFE, K- means, ReliefF and F-score for disease prediction using Support Vector Machines. The performance of these algorithms are compared by predicting various diseases. The accuracy is compared on Wisconsin Diagnostic Breast Cancer dataset. Support Vector Machine- Recursive Feature Elimination is a method where unimportant features are removed repeatedly without using weights for estimating benchmark. But this can only be used to linear kernel SVM, because in case of non linear kernel it is quite difficult to find the weight vector. K-means is an unsupervised machine learning algorithm to classify the objects on the basis of features into k groups. It finds a division where the objects within each cluster are close to each other and far from the objects in other clusters. Using SVM-RFE, can get the highest accuracy of 97% and with the K-means which is simple method comparatively assists in attaining the accuracy of 96%. Since both of them almost gives highest accuracy any one of the can be used for feature extraction in breast cancer detection.

B. Padmapriya and et.al [5]surveyed on breast cancer analysis using data mining techniques. Datamining has turned into methodology for computing applications in the domain area of medicine. A research paper by AbdelghaniBellaachia and ErhanGuven presents an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques[11]. They used Public–Use Data SEER and dataset was preprocessed which consists of 151,886 records and has 16 fields from the Database SEER. Using these dataset, three data mining techniques namely Naïve Bayes, back-propagated neural network and the C4.5 decision tree algorithm. Based on the results of conducted experiments on these algorithms they finally conclude that C4.5 algorithm gives better performance than the other 2 techniques.

JaiminiMajali and et.al[6] presents diagnosis and prognosis of cancer disease using data mining techniques system. They have used Association Rule Mining[ARM] on decision tree algorithm in this system in which they have used Frequent Pattern algorithm under Association Rule Mining

and ID3 algorithm in Decision Tree under classification. The diagonistic analysis for various applied data mining classification technique accuracy is highly acceptable and will help the medical professionals in decision making for early diagnosis and to avoid biopsy. They have used Wisconsin data set which consists of attributes which is used as input data for FP-Growth algorithm. Among data mining classifiers and soft computing approaches, decision tree is found to be the best predictor on Wisconsin data set. With this algorithm applied to data set 94% class-labels were predicted correctly.

K. Sivakami and et.al [7] proposed a survey on various classification techniques like Support Vector Machine hybrid model Decision Tree used for prediction of breast cancer. They have taken Wisconsin Breast Cancer Dataset from machine learning repository UCI with the aim of developing accurate prediction models for breast cancer using data mining techniques. After conducting the experiments they compared on the results of three classification techniques using Weka software and the comparison results indicate that the Decision Tree-Support Vector Machine are having higher prediction accuracy than Instance-based learning(IBL), Sequential Minimal Optimization(SMO), Naïve based classifiers.

Walaa Gad and et.al [8] proposes a SVM-K-means method for diagnosis of breast cancer. The method stated uses Support Vector Machine (SVM) as a classifier along with K-means clustering algorithm as a combination. Along with that many classifiers can be used many like Naïve Bayes, Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), SVM. They have combined multiple classifiers to enhance security. The proposed method is evaluating using 2 datasets: Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) obtained from Machine learning repository UCI. SVM-K means is applied on these two datasets along with many other classifiers. For evaluating the proposed algorithm they have used accuracy, precision and recall performance measures. The results shows that SVM-Kmeans gives better results compared to other approaches. It reaches upto 99.8% accuracy, 0.99 recall and 0.99 precision.

M. Nivaashini and et.al [4] presented a survey on Breast Cancer Risk Detection for Healthcare Systems using Deep Boltzmann Machine (DBM), a deep learning approach for finding an efficient set of features and Deep Neural Network (DNN) classifier is used to part the women either into benign group or malignant group. The features which are important in diagnosing like tumour features has been collected from he UCI repository. The extracted original set of features is processed and pre-trained by a deep learning algorithm called Deep Boltzmann Machine (DBM). An optimised feature set is resulted based on the detection accuracy by piling 3 hidden layers of DBM. The proposed medical diagnosis decision making system shows that the application of DBM for feature selection and DNN for classification achieves better results and performance in the prediction and classification of breast tumors. The proposed system obtains higher detection rate of 99.73%. The exertion in the proposed technique is training the complete DBM prototype is arduous through undirected links among the visible and hidden layers in the neural network architecture. Henceforth in the upcoming research, the aforesaid limitations of the DBM method be able to overwhelmed by means of Deep Belief Network that uses directed links of the neural units among the visible and hidden layers.

Muhammad Farooq and et. al [3] emphasizes the usage of convolutional neural networks (CNN). A exceptional class of deep learning algorithms is used for feature extraction from food images. The goal here is to explore the use of pre-trained CNN model for feature extraction for classification of food images into different food categories. Addition to this it aims to explore the classification acuity of extracted features from various fully connected layers of CNN. Here, SVM classifier was used for classifying food intake using features extracted from the pre-trained CNN models, to perform multi-class classification. This compares the results of the proposed approach with previously reported results on the same image datasets. The accuracies obtained for features extracted from FC6, FC7, and FC8 layers were 94.01%, 93.06%, and 89.73%, respectively. Features that are extracted from fully connected layer FC6 along with linear SVM which is obtained from the classification provides accuracy of 94.01% with an evident increase in performance of about 7%.

III. PROBLEM STATEMENT

Breast cancer is identified as most commonly found cancer and it has been one of the main reason for cause r death among the women worldwide.

Once after the Completion of initial treatment the possibilities of the far and near recurrence will be high during the first two years. Late relapse would be uncommon in breast cancer but there is a chance of recurrence. Therefore there is a need to for the prediction for the recurrence. A system is developed for the prognosis of breast cancer recurrence or relapse based on the information provided for each and every patient using various machine learning algorithms.

1) Various algorithms of machine learning have been used to prognosis the recurrence of breast cancer which would be used to compare the accuracy of each machine learning algorithms.
2) Then the algorithm which gets the highest accuracy is used to predict recurrence of breast cancer.

IV. PROPOSED SYSTEM

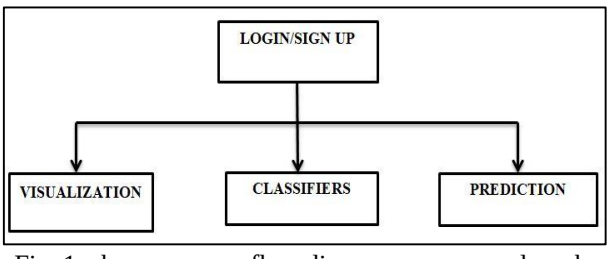


Fig. 1: shows process flow diagram or proposed work.

Breast cancer relapse prognosis contains four stages.

1) Registration Phase
2) Classification Phase
3) Visualization Phase
4) Prediction Phase

A. Registration Phase

In this phase, the user needs to enter all the required information such as user id, username, password, valid email-id, etc., required for creation of an account. After the creation of account, the user can login successfully using the authenticated username and password.

B. Classification Phase

In this phase, the model to predict the breast cancer relapse is created using training dataset for each algorithms. The algorithms used to develop the model are Logistic Regression, Deep neural network, Support Vector Machine and K-Means algorithm. The accuracy obtained by each algorithm is compared, and the algorithm which gives highest accuracy is considered for prediction phase.

C. Visualization Phase

Visualization phase represents visual output of the proposed system. The user can view the graphical representation of confusion matrix of each algorithm used and model comparison of all algorithm are done.

D. Prediction Phase

In this phase, the prediction result will be obtained that is whether the disease would be relapsed or not. The user will enter the input that is the medical details of patient for whom breast cancer recurrence should be predicted. Using the provided input the system will predict the breast cancer could be recurred or not and then the result will be displayed.

V. IMPLEMENTATION

A. Loading of Dataset

Loading of dataset is an initial stage where the datasets are loaded to the system. This dataset contains the data of the relapsed cancerous and non-relapsed cancerous patient. Based on the dataset provided the proposed system would be trained and will able to predict the recurrence. The dataset is divided as training data, testing data and validation data where 70% is used as training data, 25%is used for testing and remaining 5% is used for validation. The dataset includes attributes which are needed for prediction like time, Lymphnode_status, tumour_size, radius_mean, Concavity mean etc. Among themlymphnode status, tumour_size, time plays a very important in prediction of breast cancer relapse.

B. Data Preparation

1) Data Selection

Data selection involves selection of subset of the selected sample which must be an accurate representation of entire population.

In other words it can be explained as the selection of relevant features which would be used in model construction for algorithms.

2) Data Processing

Formatting, cleaning and sampling of data is carried out in data pre-processing step. By the above mentioned methods the data becomes more useful and informative. Formatting involves converting the data into standard format which could be understood by machine. The datasets might have missing values so in order to train the machine this has to be either removed or fixed which is carried out by cleaning process. The dataset used would be very large so this has to be reduced into sample and this sample of the selected data that would be much better and fast for inquiring and modeling solutions even before considering the whole dataset.

3) Data Transformation

This step involves formation of features from large amount of data. The three main procedures used here are scaling, decomposition and aggregation. The preprocessed data may contain different scales .Using scaling process the data would be scaled to particular range. In order to make the process of making machine to learn it is better to divide large concept into smaller one. This will be carried by decomposition, even the vice versa will be possible where the constituent parts are aggregated to form complex concepts using aggregation method in order to make machine to understand.

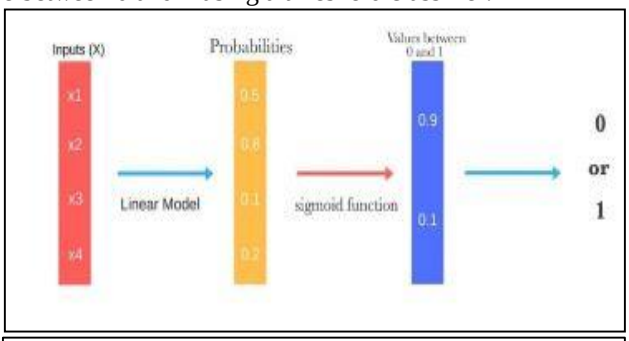
C. Classification

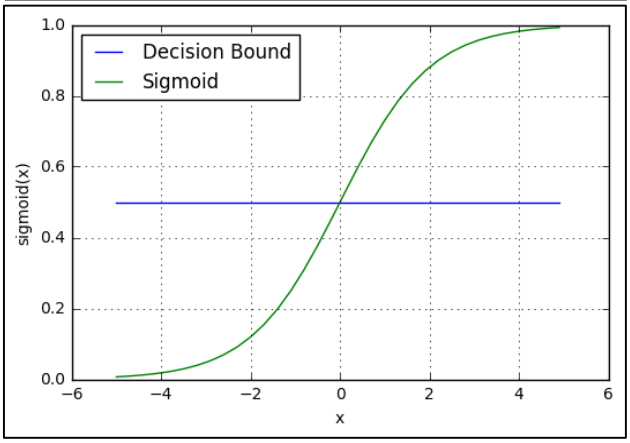
1) Logistic Regression

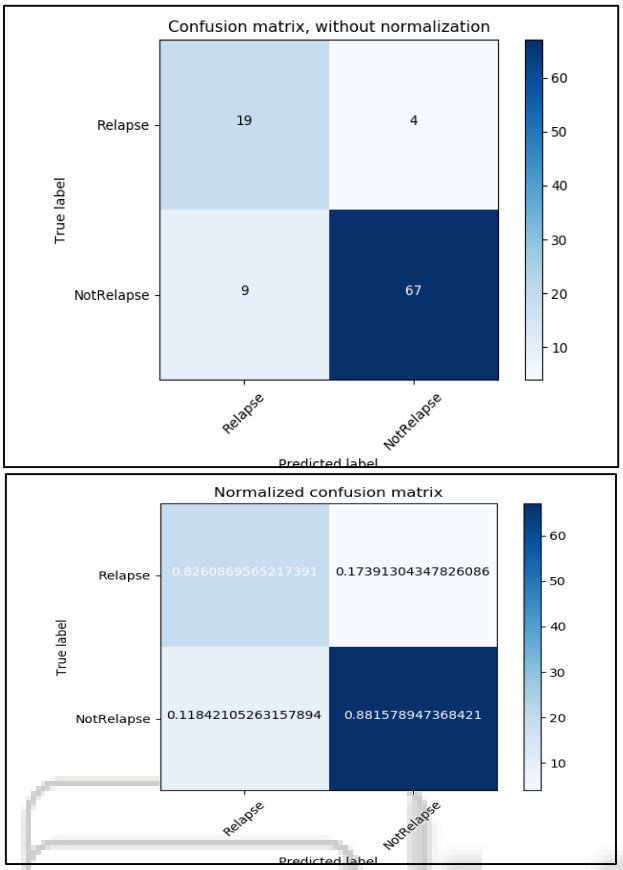
Logistic Regression algorithm is one of classification algorithm of machine learning which is used for classifying discrete set of classes. Logistic regression algorithm uses sigmoid function to transform the output into probability value which is useful for classifying two or more discrete classes.

Logistic Regression indicates the relativeness between the independent variable and an one or more dependent variables, by calculating the probabilities using the logistic function used.

Sigmoid function is also used to transform these probability values into binary values as it is helpful in making prediction. The S-shaped sigmoid curve will be has plotted using any real-valued number. This will be matched it into a value within the range of 0 and 1 but not exactly at that points. These values will be computed and the computed value would be between 0 and 1 using a threshold classifier.





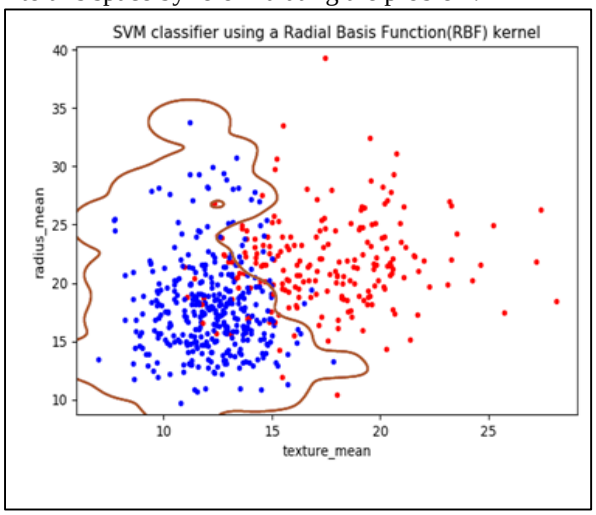


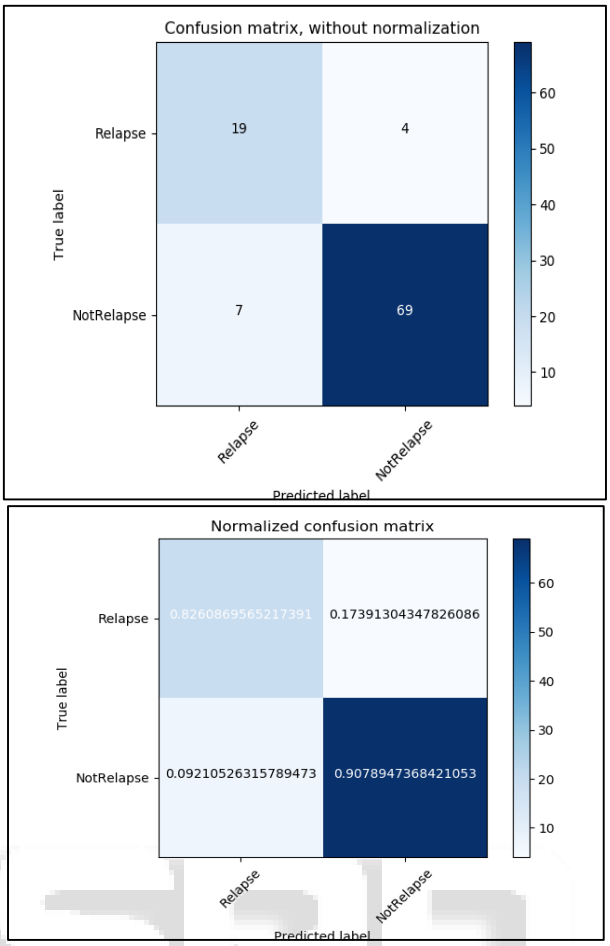
2) Support Vector Machine

A Support Vector Machine(SVM) is an algorithm in machine learning used for grouping purposes. This process involves finding a hyperplane and this hyperplane is used to draw the margin between the two classes which is useful for prediction. The support vectors are the vectors that define hyperplane.

a) Algorithm:

1) Optimal hyper plane must be defined: maximize margin.
2) Linear separable problems are defined for optimal hyper plane which has a penalty term for misclassifications.
3) The information must be mapped to large dimensional space which helps in performing classification in linear decision surface. The data should be mapped implicitly to this space by reformulating the problem.





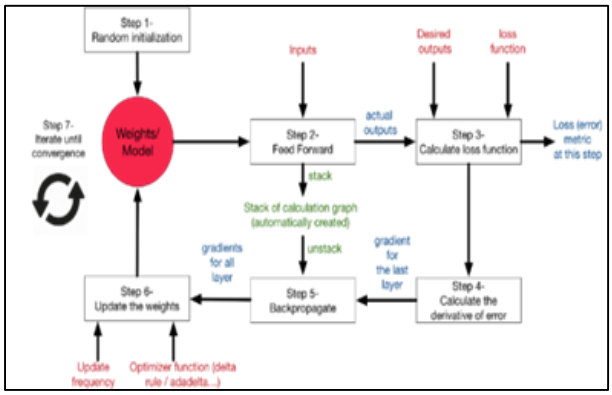
3) Deep Neural Network

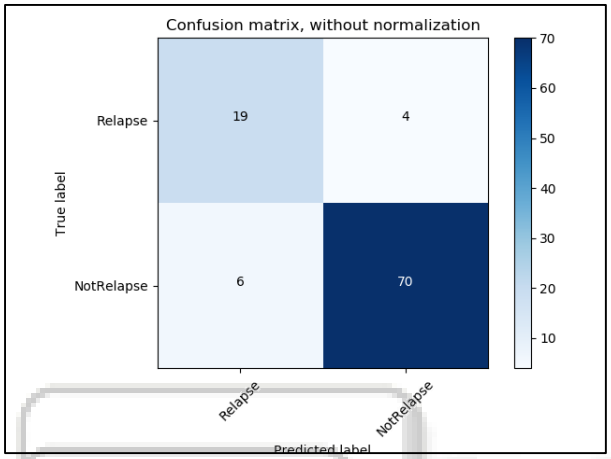
A Deep Neural Network is nothing but artificial neural network containing at least two hidden layer. A neural network with more than two layers has a certain level of complexity. More complexity of data is captured with the use of more number of hidden layers. Neural network is a technology to imitate the job of the human brain.

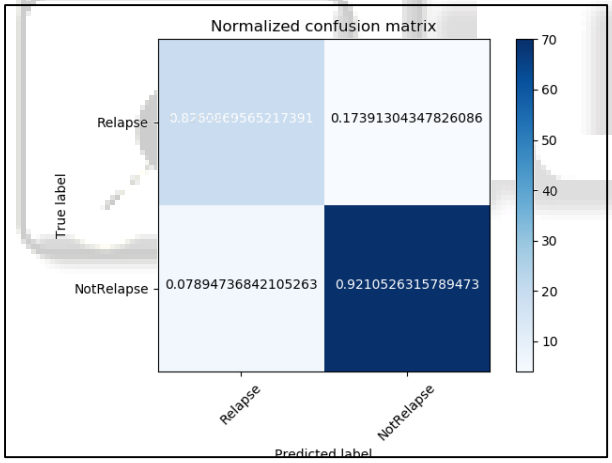
Deep neural network is a network which consists of three layers, an input layer, output layer and many hidden layers but it can also have many hidden layers. Each layer carry out particular kinds of sorting and sequencing in a process which refer to as feature hierarchy.

It represents the particular kind of machine learning which is used to group and sequence the information in ways that go beyond simple input/output protocols.

Back propagation algorithm is an specific method for increasing the correctness of prediction of machine learning and data mining. Artificial neural networks use back propagation which is a precise rule to calculate a gradient descent with respect to weights. The system outputs achieved are compared with the required outputs from the algorithm. On observing this system are tuned to decrease the difference between the two outputs as far as possible. Since the weights will get improve in the backwards direction from output layer towards input layer, this algorithm is called as back propagation algorithm.







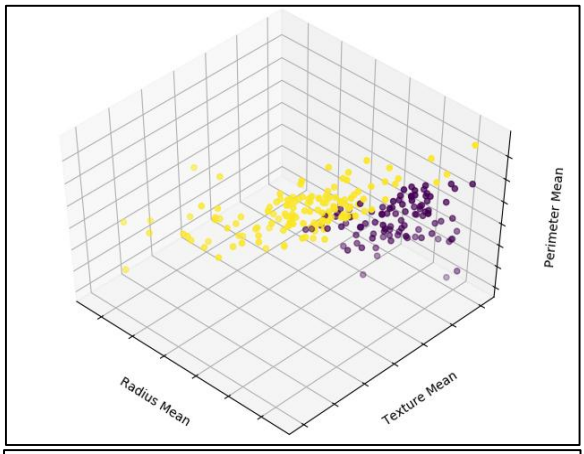
4) K-Means Clustering

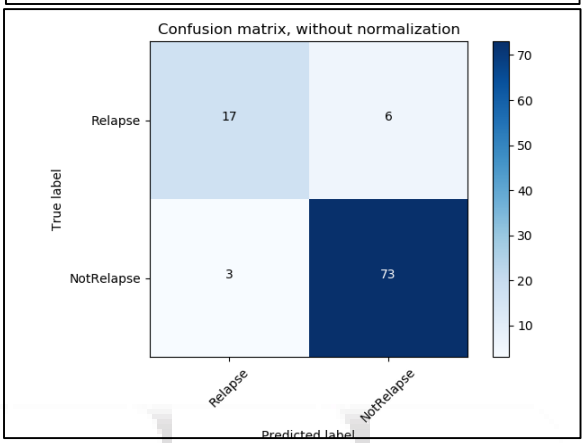
It is an autonomously algorithm which categorize the objects into k clumps of likeness. For calculating the similarity, euclidean distance is used as calculation.

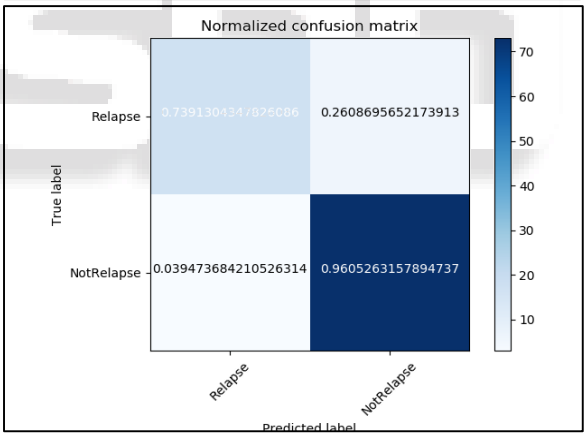
Algorithm:

1) Initialize points k called as means randomly.
2) Each item should be categorized to its closest mean and the mean's coordinates value are updated, these are the means of the objects that is grouped.
3) Reiterate the process for the given number of times and in the end the clumps are obtained.

The points specified above contain the mean values of the items which are categorized in it and hence they are called as means. An automatic method is used to assign initial value to the average at random objects in the dataset. Different method is used to assign initial value to the averages at arbitrary values between the borders of the dataset.

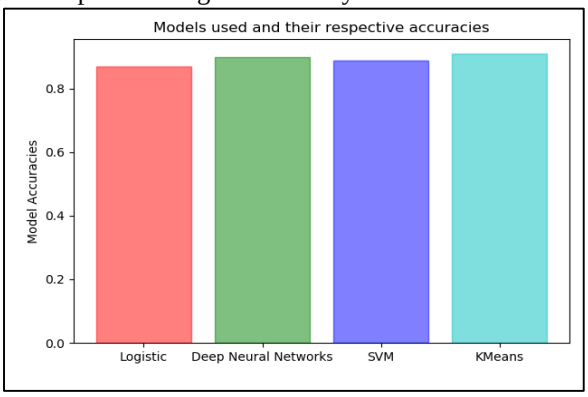






5) Model Comparison

The below figure depicts the model used and respective accuracy. The model with the highest accuracy is used for predicting breast cancer relapse. KMeans clustering algorithm provides highest accuracy.



VI. CONCLUSION

The samples of breast cancer collected from the clinics will be used to classify whether can be recurred or not. It is important to choose right amount of features as it is an important in increasing the accuracy of classification algorithm. Choosing right amount of feature involves removing unwanted attributes. This process is called dimensionality reduction and it helps in increasing the speed of classification process. This process eventually aids in increasing overall performance of machine learning algorithm. The implementation of classification algorithms gives better results. Hence various classification algorithm could be used to reach better results and performance in breast cancer relapse prognosis.

REFERENCES

[1] Wei Wang, Min Zhang, Dan Wang and Yu Jiang, "Kernel PCA feature extraction and the SVM classification algorithm for multiple-status, through-wall, human being detection" EURASIP JournalonWirelessCommunicationsandNetworking,Sept .2017.
[2] SandeepKaur, Dr. SheetalKalra, " Feature Extraction Techniques Using Support Vector Machines In Disease Prediction" International Journal of Advance Research inScienceandEngineering,Vol.No.5,IssueNo.05,May2016.
[3] Muhammad Farooq and Edward Sazonov, "Feature Extraction Using Deep Learning for Food Type Recognition" Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa AL 35487,USA.
[4] M.Nivaashini and R.S.Soundariya, "Deep Boltzmann Machine based Breast Cancer Risk Detection for Healthcare Systems" International Journal of Pure and Applied Mathematics Volume 119 No. 7 2018, 581-590.
[5] B.Padmapriya, T.Velmurugan, "A Survey on Breast Cancer Analysis Using Data Mining Techniques, " Computational Intelligence and Computing Research,2015.
[6] J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe, "Data mining techniques for diagnosis and prognosis of cancer," Int. J. Advanced Research in Computer and Communication Engineering, vol. 4, pp. 613– 616,2015.
[7] SivakamiK,"Mining big data: breast cancer prediction using DT – SVM hybrid model,".
[8] Gad W, "SVM-Kmeans: Support Vector Machine based on Kmeans clustering for breast cancer diagnosis"
[9] SujoyMondal, SoumadipGhosh, BhaskarGhosh, "A Comparative Study of Breast Cancer Detection Based on SVM and MLP BPN Classifier" Academy of Technology, Hooghly, West Bengal, India.
[10] Rajat K, GoutamSaha, SudipMandal, "A Comparative Study on Disease Classification using Different Soft Computing Techniques"

