BREAST CANCER DETECTION USING RANDOM FOREST, KNN AND SVM

Rishabh Khosla
Maharaja Agrasen Institute of Technology, New Delhi, India.

ABSTRACT
Breast Cancer is the most often identified cancer among women and major reason for increasing mortality rate among women. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. Data mining techniques contribute a lot in the development of such system. For the classification of benign and malignant tumour I have used classification techniques of machine learning in which the machine learns from the past data and can predict the category of new input. This is a relative study on the implementation of models using Random Forest, K nearest neighbours and SVM. This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this machine learning classification methods have been used to fit a function that can predict the discrete class of new input.

I. INTRODUCTION

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci: 1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability. In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. This study is on the first case. The study will include finding the features that contribute most to the prediction variable, this is known as feature selection. Various Machine Learning algorithms including Random Forest, KNN and SVM are going to be used to determine whether the tumor is benign or malignant.

Machine Learning Algorithms
Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Machine Learning Algorithms used are:
1.Random Forests
2.KNN
3.SVM

II. LITERATURE REVIEW

Several researchers studied on predicting earlier way of breast cancer diagnosis. The learning systems is of data mining techniques, machine learning techniques and the hybrid form of data mining and machine learning systems. The following algorithms are widely used in breast cancer prediction:

1. Decision Trees
2. Artificial Neural Network
3. Genetic algorithms
4. Support Vector Machines

Decision Trees: The best classification algorithms widely used in medical applications is the decision trees. It is in form of graph-based systems. The eminent decision trees algorithms were Quinlan's, ID3, C4.5 and C5

Genetic Algorithms: Genetic algorithms are dynamic heuristic techniques. A genetic fitness function is calculated to find the genetic approach. In utilized the genetic systems for the forecast of breast tumor. This system is hybrid with the decision tree, ANN and logistic 4 regression. They used 699 records acquired from the breast cancerous patients at the University of Wisconsin. They utilized 9 indicator variables and 1 result variable for the information investigation with 10-fold cross approval. The researchers asserted that their genetic prediction model gives precision as much as 99%.

Support Vector Machines: Support Vector Machines belongs to the class of supervised learning systems. It is one of the best optimization procedures. This reduces the over-flowing of the trained data. The goal is to find the optimized decision boundaries to predict the breast cancer at the earlier stage. C-Support Vector Classification Filter (C-SVCF) algorithms was used to spot and kill exceptions in breast malignancy survivability information sets.
Consequences of their methodology demonstrated performance enhancement of breast tumor survivability expectation models by enhancing information quality. This script 5 unmistakably draws consideration towards the utilization of SVM for anticipating better survival rates.

III. ML MODEL PHASES

The stages are as follows:
Data Pre-processing: Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Data Visualization:

Data visualization is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed. Data visualization is regularly used as an avenue to monetize data as a product. There are five basic plots in data visualization. They are:

- Line Plot
- Bar Chart
- Histogram Plot
- Box and Whisker Plot
- Scatter Plot

Feature Selection:

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

- simplification of models to make them easier to interpret by researchers/users,
- shorter training times,
- to avoid the curse of dimensionality,
- enhanced generalization by reducing overfitting (formally, reduction of variance)

The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information. Redundant and irrelevant are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

Training and Testing the model Training Data

The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.

Test Data

The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the
training set or has simply memorized it

Evaluation of Accuracy

Accuracy is calculated with the following formula − ACC = (TP + TN)/(TP + TN + FP + FN) Where, TP is the number of true positives TN is the number of true negatives FP is the number of false positives FN is the number of false negatives. Precision is the fraction of the tumors that were predicted to be malignant that are actually malignant. Precision is calculated with the following formula − PREC = TP/(TP + FP)

Recall is the fraction of malignant tumors that the system identified. Recall is calculated with the
following formula − R = TP/(TP + FN)

IV. DIFFERENT ML MODELS USED

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: 11 For b = 1, ..., B: 1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb. 2. Train a classification or regression tree fb on Xb, Yb.

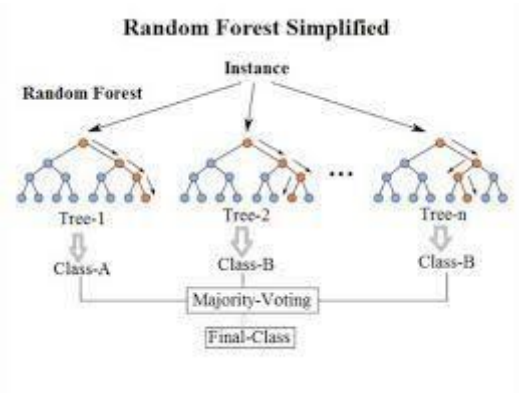

Fig 1: Random Forest Simplified

KNN

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a nonparametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether kNN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. k-NN is a type of instancebased learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

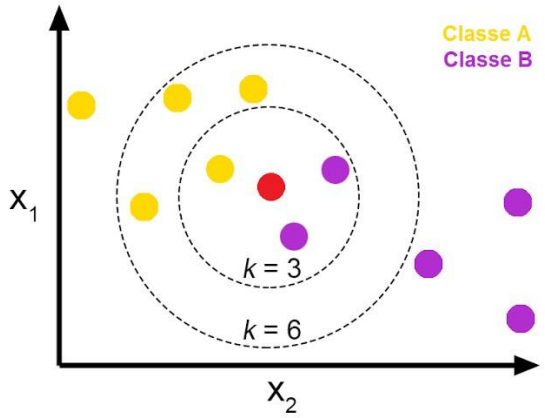

Fig 2: K NEAREST NEIGHBOR

SUPPORT VECTOR MACHINE

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a nonprobabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. In addition to performing linear classification, SVMs can efficiently perform a nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

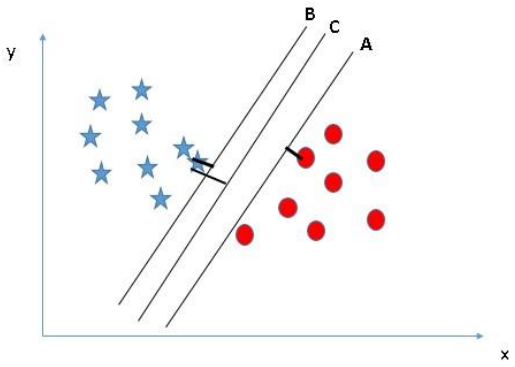

Fig 3: SUPPORT VECTOR MACHINE

V. RESULTS

On running **Random Forrest** on the dataset after appropriate feature selection, following were the results:
The number of True Negatives were 107.
The number of True Positives were 60.
The number of False Negatives were 3.
The number of False Positives were 1. The accuracy calculated from the confusion matrix was 97.6%.

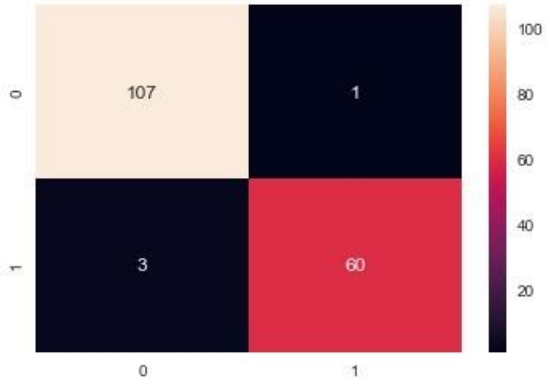

Fig 4: Confusion Matrix for Random Forest.

On running **K Nearest Neighbors** on the dataset after appropriate feature selection, following were the results:
The number of True Negatives were 107.
The number of True Positives were 49.
The number of False Negatives were 14.
The number of False Positives were 1. The accuracy calculated from the confusion matrix was 91.2%.

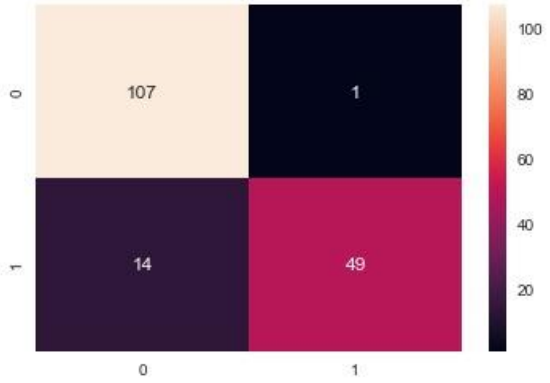

Fig 5: Confusion Matrix for KNN.

On running **Support Vector Machine** on the dataset after appropriate feature selection, following were the results:
The number of True Negatives were 104.
The number of True Positives were 4.
The number of False Negatives were 59.
The number of False Positives were 5. The accuracy calculated from the confusion matrix was 62.5%.

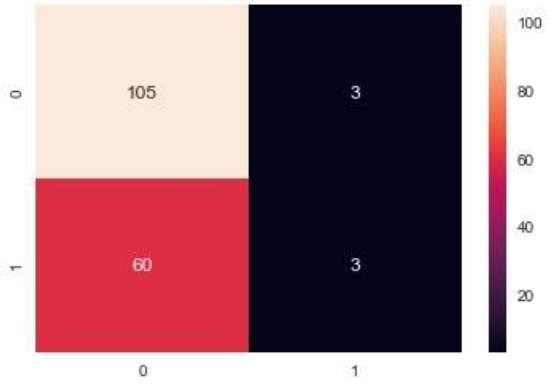

Fig 6: Confusion Matrix for SVM.

VI. CONCLUSIONS

There are different data mining techniques that can be used for the prediction of breast cancer to predict whether the

tumor is benign or malignant. In this study Random forest, K Nearest Neighbors and SVM were used and the accuracy for the prediction on the dataset namely, Breast Cancer Wisconsin (Diagnostic) Data Set was found to be 97.6%, the highest of the three, using Random Forest. Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

VII. REFERENCES

[1]http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

[2]https://arxiv.org/abs/1609.04802

[3]http://torch.ch/blog/2016/02/04/resnets.html [4]Pydata.org

[5] Campus.datacamp.com

[6] Saeed Al-Mansoori, Intelligent Handwritten Digit Recognition using Artificial Neural Network, Int.
Journal of Engineering Research and Applications , Vol. 5, Issue 5, (Part -3) May 2015, pp.46-51

[7] Sonali B. Maind, Priyanka Wankar, Research Paper on Basic of Artificial Neural Network,
International Journal on Recent and Innovation Trends in Computing and Communication, Volume:
2 Issue: 1
96 – 100

[8] J. Cao, M. Ahmadi and M. Shridar, "A Hierarchical Neural Network Architecture For Handwritten
Numeral Recognition", Pattern Recognition, vol. 30, (1997)