International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Prediction of Breast Cancer Using Ensemble Learning

Kranti Avhad[1], Gurusiddha Kore[2], Rishi Kaul[3], Sourav Jethwa[4], Pooja Mundhe[5]

[1,2,3,4] *B.E. Student, Department of Information Technology, MIT College of Engineering, Pune, Maharashtra, India*
[5] *Professor, Department of Information Technology, MIT College of Engineering, Pune, Maharashtra, India*

Abstract - *Breast cancer is the most common type of cancer in females in the United States. More 246,660 cases of invasive breast cancer will be diagnosed in females in the United States in 2016. Early diagnosis and successful rate led to a decrease of 37 percent mortality (deaths) from breast cancer between 1990 and 2013. Hence, having a system that would allow early detection and prevention would be very useful, which would increase breast cancer survival rates. Studies on cancer have recently tried to link machine learning to prediction and prognosis for cancer. Machine learning often draws from statistics and probability but it is much more efficient as it allows to make inferences or decisions that otherwise cannot be taken using traditional statistical methodologies. We have used machine learning classification techniques to classify benign malignant tumors, in which machine learns from past data and predicts the new input category. This paper is a relative study of implementation of models using Support Vector Machine (SVM), k-Nearest Neighbor (k-NN) and Decision tree is done on the dataset. With respect to the results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm is measured and compared. All these models are integrated with the help of ensemble learning. This improves the performance of model.*

Key Words: **Breast Cancer Prediction, Benign, Malignant, Ensemble Learning, Classification, Overfitting, SVM, k-NN, Decision Tree**

1. INTRODUCTION

Breast Cancer is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. According to the statistics provided by the World Cancer Research Fund, it is estimated that more than 2 million cases have been recorded, out of which approximately 626,679 deaths have been reported. In new cases of cancer, breast cancer constitutes 11.6 percent of all cancers and results in 24.2 percent of women's cancer. Usually people visit doctor immediately in case of any sign or symptom who may refer to an oncologist, if necessary.

1.1 Imaging Tests

Mammogram, Magnetic resonance imaging (MRI) of breast, Ultrasound of breast, X-ray of the breast, Tissue biopsy: Removal of the tissue of the breast for examination by a pathologist. Sentinel node biopsy: Once breast cancer is confirmed, patients regularly undergo sentinel node biopsy. This helps the detection of cancer cells in lymph nodes to confirm breast cancer metastasis into the lymph system. Oncologists may also recommend additional tests and procedures, if necessary. Some examinations and procedures are carried out in the traditional way in which the breast cancer is diagnosed. These tests include Breast exam Mammogram Breast ultrasound Biopsy. We can also use Machine learning techniques as an alternative for classifying benign and malignant tumors. This approach improves the prediction, so that patients can be offered appropriate care at the right time and thereby increases the survival rate. Machine learning approaches are being used today in a variety of medical applications including the identification and classification of tumors. The prediction of breast cancer has long been seen in the medical and healthcare communities as an important research problem. [1]

There are various types of breast cancer, with various stages or spread, aggressiveness and genetic makeup. A variety of statistical and machine learning methods have been employed to build various predictive models to detect breast cancer. . In this paper, we have used three algorithms – SVM, K-Means and Decision Tree for prediction. We are going to integrate all these algorithms by ensemble [2] method to get higher accuracy.

1.2 Breast Cancer: An Overview

Breast Cancer is the most common type of cancer in women worldwide. It is also the main cause of death from cancer among women globally. Cancer starts when cells begin to grow out of control. Breast Cancer cells usually form a tumor that can often be seen on an x-ray or as a lump. [1]

Breast Cancer lumps are benign or non-cancerous and malignant or cancerous. Non-cancerous breast tumors are abnormal growths, it does not spread. Any cancer lump must be checked by the health professional to determine whether it is benign or malignant and if it might cause risk in the future.

1.2.1 Breast Cancer Types

Three types of breast tumors may be classified as benign breast tumors, in situ cancers and invasive cancers. Most mammogram-detected breast tumors are benign. They are

non-cancerous growths, and cannot spread to other organs. In certain cases the differentiation between such benign masses and malignant mammography lesions is difficult. The cancer is called in situ or non-invasive if the malignant cells have not passed through the basal membrane but are completely contained in the lobule and ducts. The cancer is invasive if it has penetrated through the basal membrane and spread into the surrounding tissue. So early breast cancer detection is important. In our study, we focus on the distinction between benign and malignant tumors. The goal of these predictions is to assign patients to either a non-cancerous "benign" group or a cancerous "malignant" group.

2. METHODOLOGY

We obtained the breast cancer dataset from kaggle repository and used spyder as the platform for the purpose of coding. Our methodology involves use of classification techniques like Support Vector Machine (SVM), k-Nearest Neighbor and Decision Tree with Dimensionality Reduction technique.

2.1 Data Splitting

Data set contains certain number of instances. We will need to split these instances into training and testing sets. These are split with the ratio of 80% to 20% for training and testing purpose respectively.

2.3 Dimensionality Reduction

Dimensionality reduction is a process in which the number of independent variables is reduced to a set of principle variables by removing those that are less important in predicting the result. It is used to obtain two dimensional data such that machine learning models can be properly visualised by plotting the prediction regions and prediction boundaries for each model. Whatever the number independent variables, we often end up with two independent variables by applying the appropriate technique of reducing dimensionality. There are two processes namely, Feature Selection and Feature Extraction.

2.3.1 Feature Selection

Feature selection is the selection process of a subset of specific features (variables, predictors) to be used in model construction. It helps to simply the model to make it easier for researchers to predict the model. It reduces overfitting and enhances generalisation.

2.3.2 Model Selection

We have chosen three different types of classification algorithms in machine learning.

1) Support Vector Machine (SVM)
2) k-Nearest Neighbor (k-NN)
3) Decision Tree

2.3.3 Support Vector Machine (SVM)

A support vector machine (SVM) is a supervised model of machine learning, which uses classification algorithms for binary classification. Support vector machines are today's most efficient predictive accuracy classification algorithm. The methods in the SVM algorithm are based on strong mathematical principles and the theory of statistical learning. SVM does well in problems of pattern recognition and is used as a training algorithm to study data classification and regression rules. SVM is mostly used when number of features and number of instances are high. A binary classifier can be built with the help of SVM. The binary classifier is designed using a hyper plane where it is a line in more than 3 dimensions. The hyper plane performs the task of dividing the members into one of the groups.

Hyper plane of SVM is built on mathematical equations. The hyper plane is WTX = 0, which is identical to the line equation y = ax+b. Here W and X represent vectors where the vector W is always normal to the hyper plane. WTX represents the dot product of vectors. Since SVM deals with the dataset when the number of features is greater, in this case we need to use the equation WTX = 0 instead of using the line equation y = ax+b.

If a set of training data is given to the machine, each data item will be assigned to one or the other categorical variables, a SVM training algorithm builds a model that plots new data item to one or the other category. In an SVM model, each data item is represented as points in an n dimensional space where n is the number of features where each feature is represented as the value of a particular coordinate in the n-dimensional space. Classification is carried out by finding a hyper plane that divides the two classes proficiently. Later, new data item is mapped into the same space and its category is predicted based on the side of the hyper plane they turn up.

2.3.4 k-Nearest Neighbor (k-NN)

k- Nearest Neighbor is a supervised learning algorithm as the data given to it is labelled. It is a nonparametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset. [3] It is employed in solving both classification and regression tasks. In classification technique, it classifies objects based on the k-closest training examples in the feature space.

The working principle behind KNN is it presumes that alike data points lie in the same surroundings. It reduces the burden of building a model, adapting a number of parameters, or building furthermore assumptions. It catches the idea of proximity based on mathematical formula called as Euclidean distance, calculation of distance between two points in a plane. Suppose the two points in a plane are A(x0,y0) and B(x1,y1) then the Euclidean distance between them is calculated as follows

$$\sqrt{(x0-x1)^2 + (y0-y1)^2}$$

An object to be classified is allotted to the respective class which represents the greater number of its nearest neighbors. If k takes the value as 1, then the data point is classified into the category that contains only one nearest neighbor. Given a new input data point, the distances between that points to all the data points in the training dataset are computed. Based on the distances, the training set data points with shorter distances from the test data point are considered as the nearest neighbors of our test data. Finally, the test data point is classified to one of the classes of its nearest neighbor. Thus the classification of the test data point hinges on the classification of its nearest neighbors.

Choosing the value of K is the crucial step in the implementation of KNN algorithm. The value of K is not fixed and it varies for every dataset, depending on the type of the dataset. If the value of K is less the stability of the prediction is less. In the same manner if we increase its value the ambiguity is reduced, leads to smoother boundaries and increases stability. In KNN, assigning a new data point to a category entirely depends upon K's value. K represents the number of nearest training data points in the proximity of a given test data point and then the test data point is allotted to the class containing highest number of nearest neighbors(i.e. class with high frequency).

2.3.5 Decision Tree

Decision Tree is a supervised learning algorithm. It is used to solve problems regarding regression and classification. By learning the rules of judgement derived from the training data, the decision tree is used to construct a training to predict the class or value of target variables. By using tree representation the decision tree algorithm attempts to solve the problem. The internal node of the tree corresponds to an attribute, and the leaf node corresponds to a class label.

Decision tree creates a tree structure in the form of classification or regression models. It breaks up a data into smaller and smaller subsets, while creating a related decision tree incrementally at the same time. The end result is a tree with decision nodes and leaf nodes. Decision node has two or more branches. Leaf node is a grouping or decision. The top decision node in a tree that corresponds to the best predictor called root node. Decision trees are capable of handling both categorical and numerical data. Leaf node is a grouping or decision.

The top decision node in a tree that corresponds to the best predictor called root node. Decision trees are capable of handling both categorical and numerical data.

To build a decision tree, the first step is splitting. In this the datasets are partitioned into subsets. After splitting of datasets, pruning is done. The shortening of branches of the tree is known as pruning. It is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch. Pruning is useful as classification trees may fit the training data well, but may do a poor job of classifying new values. A simpler tree often avoids over-fitting. The final step is tree selection. It is the process of finding the smallest tree that fits the data. This tree yields the lowest cross-validated error.

3. RESULTS AND DISCUSSIONS

As our dataset contains 32 attributes dimensionality reduction leads greatly to the reduction of multidimensional data to a few dimensions. Of all the three applied algorithms Support Vector Machine, k Nearest Neighbor and Decision Tree, SVM gives the highest accuracy of 99% when compared to other two algorithms. So, we propose that to predict Breast Cancer Occurrence with complex datasets, SVM is the best suited algorithm.

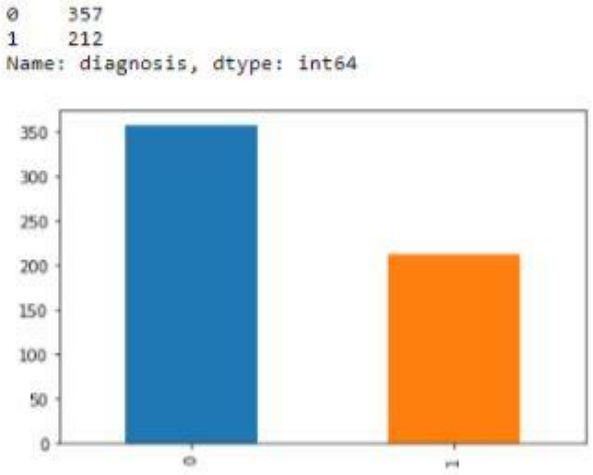

Fig-1: No. of Benign and Malignant instances

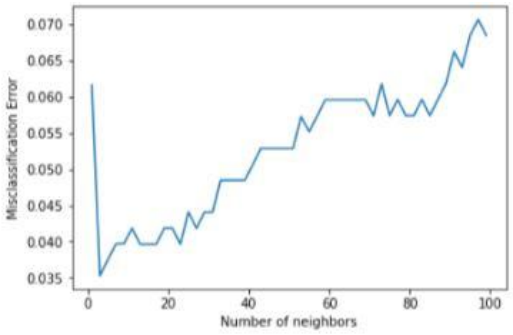

Fig-2: k-NN Classification

Table 1: Comparison of accuracies of various algorithms

Sr. no	Algorithm	Accuracy
1	Support Vector Machine	99%
2	K-Nearest Neighbour	98.02%
3	Decision Tree	93%

ACKNOWLEDGMENT

We would like to express our sincere appreciation to Prof. Pooja Mundhe (MITCOE, Pune) for her constructive and valuable suggestions. Her ideas and insight proved helpful throughout the process.

REFERENCES

[1] Ch. Shravya, K. Pravalika, Shaik Subhani, 'Prediction of Breast Cancer Using Supervised Machine Learning Techniques', *International Journal of Innovative Technology and Exploring Engineering(IJITEE), Volume-8 Issue-6,* 2019.

[2] Naresh Khuriwal, Nidhi Mishra Department of Computer Engineering Poornima University Jaipur, India, 'Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm', *IEEMA Engineer Infinite Conference (eTechNxT)*, 2018.

[3] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan Department of Computer Science and Engineering Khulna University of Engineering & Technology Khulna-9203, Bangladesh, 'Prediction of Breast Cancer Using Support Vector Machine and K-NN Neighbors', *IEEE Region 10 Humanitarian Technology Conference (R10-HTC),* 2017

[4] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, 'Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm', *2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool,* 2016.

[5] Mohamed Bahaj Department of Mathematics and Computer Science Faculty of Sciences and Techniques, Hassan 1st University Settat, Morocco, 'Feature selection with Fast Correlation-Based Filter for Breast cancer prediction and Classification using Machine Learning Algorithms', *International Symposium on Advanced Electrical and Communication Technologies(ISAECT),* 2018.

BIOGRAPHIES

Kranti Avhad
Pursuing B.E. in
Information Technology
MITCOE, Pune.

Gurusiddha Kore
Pursuing B.E. in
Information Technology
MITCOE, Pune.

Rishi Kaul
Pursuing B.E. in
Information Technology
MITCOE, Pune.

Sourav Jethwa
Pursuing B.E. in
Information Technology
MITCOE, Pune.

Pooja Mundhe
Professor
Department of I.T.
MITCOE, Pune