Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 3, pp. 383–392
DOI : 10.47974/JIOS-1348

Breast cancer prediction using supervised machine learning techniques

Pankaj Dadheech *
Department of Computer Science & Engineering
Swami Keshvanand Institute of Technology, Management & Gramothan
Jaipur
Rajasthan
India

Vijay Kalmani †
Department of Computer Science & Engineering
Jain College of Engineering
Belgaum
Karnataka
India

Sanwta Ram Dogiwal §
Department of Information Technology
Swami Keshvanand Institute of Technology, Management & Gramothan
Jaipur
Rajasthan
India

Vijay Kumar Sharma ‡
Department of Computer and Communication Engineering
Manipal University Jaipur
Jaipur
Rajasthan
India

* *E-mail:* **pankajdadheech777@gmail.com (Corresponding Author)**
† *E-mail:* **vijaykalmani@hotmail.com**
§ *E-mail:* **dogiwal@gmail.com**
‡ *E-mail:* **vijaymayankmudgal2008@gmail.com**

© Taru Publications

Ankit Kumar [@]
Saroj Kumar Pandey [#]
Department of Computer Engineering & Applications
GLA University
Mathura
Uttar Pradesh
India

Abstract

Breast cancer is one of the most prevalent diseases in India's urban regions and the second most common in the country's rural parts. In India, a woman is diagnosed with breast cancer growth every four minutes, and a woman dies from breast cancer sickness every thirteen minutes. Over half of breast cancer patients in India are diagnosed with stage 3 or 4 illness, which has extremely low survival rates; hence, an urgent need exists for a rapid detection strategy. To forecast if a patient is at risk for breast cancer, we utilise the classification techniques of machine learning, in which the machine learning model learns from the previous information and can anticipate on the new information that is generated by the data. To create a model using Logistic Regression, Support Vector Machines, and Random Forests, this dataset was collected from the UCI repository and studied in this study. The primary goal is to improve the accuracy, precision, and sensitivity of all the algorithms that are used to categorise data in terms of the competency and viability of each and every algorithm. Random Forest has been shown to be the most accurate in classifying breast cancer, with a precision of 98.60 percent in tests. The Scientific Python Development Environment is used to carry out this machine learning study, which is written in the python programming language.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Breast cancer prediction, Logistic regression, Supervised machine learning, Support vector machine.*

1. Introduction

In the year 2012, counted 70,218 women died of breast cancer in India, the highest number in the world for that year. India has a predominantly young population and hence the number of ladies being determined to have breast cancer is just going to increment. Cancer can be described as a class of infectious disease which portrays as an abnormal tissue (malignant) growth. Breast cancer growth is an illness wherein threatening cells structure in the tissues of the breast. Breast cancer tumors can be classified

[@] *E-mail:* **kumar.ankit@gla.ac.in**
[#] *E-mail:* **saroj.pandey@gla.ac.in**

into two kinds, one is non-cancerous, which are 'Benign', and the other is cancerous, which are 'malignant'. Benign tumors are some of the times not harmful to the body, so when a benign tumor is analyzed, the tissue is left without any diagnosis as it is not harmful. On the other hand, malignant tumors are much harmful and forceful in light of the fact that the cancerous and it can harm any surrounding tissue, so when a tumor is found to be malignant, the specialist will perform a biopsy to determine the grimness or fierceness of the tumor.

In a conventional way, breast cancer can be diagnosed through numerous tests, including a mammogram, ultrasound, MRI, biopsy, Lab Tests. Whenever required the oncologist (i.e., Cancer Specialist) may likewise add some other tests or procedures. The pathologist is one who examines and elimination the tissue of the breast. When breast cancer is confirmed, patients frequently encounter biopsy which assists to identify malignant cells in lymph nodes. All this method and procedures are time-consuming, so as an option we can analyze benign and malignant tumors by employing Machine Learning and data mining techniques [1]. Utilizing Machine Learning we can earlier diagnose the Breast Cancer which can improve the prediction and endurance rate with the goal that patients can make the best clinical treatment at the right time.

This article discusses a study on the extraction of breast cancer features using Machine Learning algorithms. Testing three models—linear regression, support vector machine (SVM), and random forest—is the subject of this investigation. Breast cancer can be correctly diagnosed if a patient is appropriately classified into the benign and malignant classes and relevant research is conducted. Predicting breast cancer can benefit from ML's improved ability to uncover simple highlights from large datasets. Machine learning classification approaches will be extremely beneficial in classifying benign and malignant cancers [1].

2. Related Work

The fundamental investigation which was just centered around the headway of the predictive techniques to accomplish great precision directed AI strategies can be utilized for foreseeing a legitimate result of the sickness inspiration or not. Their examination connoted that the reconciliation of multidimensional information alongside various grouping strategies include determination, and dimensionality decrease methods can give a propitious instrument to deduction right now, thought was proposed by Ch. Shravya in their investigation [1]. There is a great

deal of clinical information, different information mining and machine learning strategies are accessible. A significant test that was looked at by HibaAsri et al in their examination was to manufacture accurate and computationally productive classifiers for new clinical applications. There were four significant calculations that they utilized for this reason which are k-NN, SVM, Naïve Bayes and C4.5 on the dataset which was taken from Wisconsin Breast Cancer. They attempted to look at productivity and adequacy both regarding its exactness, accuracy, affectability, and explicitness for this reason SVM which they got about 97.13% and it beat every single other calculation [2]. An age where big data and data mining tasks suffer from plenty of challenges. There were some ideas which were proposed in this study done by Haifeng Wang et al [3], it helped in the detection of cancer and they used four different data mining models as well there were eight different hybrid models which were tested using two data sets. PCA, as a dimension reduction technique has taken out some advantages in terms of the accuracy and efficiency for the prediction that whether the person has cancer or not. They also proposed the idea that other than linear reduction methods there are many non-linear reduction methods which can be tested such as k-means [3].

3. Methodology

In this experiment, we take the Breast Cancer datasets (original) from the UCI Repository. Breast Cancer UCI repository Dataset has 699 instances (Benign (B): 458, Malignant (M): 241), two labels in the dataset as (65.5%: M and 34.5%: B), and 32 integer valued attributes. Our methodology involves the compare the machine learning classification technique like Random Forest, Support Vector Machine and Logistic Regression with dimensionality reduction technique for example Principal Component Analysis (PCA) and Feature Engineering.

3.1 *Dimensionality Reduction*

In this method, the end result is generated by decreasing the number of the independent variables to a set of principle variables by eliminating those variables that are less important. Many machine learning issues involve thousands or even a huge number of features for each training instance not only do they make trainers extremely slow it can also make it much harder to find a good solution and it additionally leads to the overfitting of the model. In this way, Dimensionality reduction is a very

crucial step considered while creating any model. So, it generally achieves by two methods- Feature Extraction and Feature Selection.

3.2 *Feature Extraction*

In this, a valuable feature is plucked from the original feature dataset and the high dimensional space data is converted into a low dimensional space data. This transformation is sometimes irreversible and many times useful information is lost. In this research, the dataset has 32 features so it's multidimensional data, which are associated with cell parameters. In this way, the most specific features are not given by this. So, we have to used different technique which is feature projection technique i.e., PCA (Principal Component Analysis).

3.3 *Feature Selection*

Using different methodologies based on their accuracy, prediction mistakes, or other information, we can glean a portion of the whole feature set. Feature selection can be used to prevent feature duplication or to restrict the amount of features in order to avoid overfitting the form.

3.4 *Model Selection*

This phase helps in achieving the best result of prediction therefore it is also known as Algorithm Selection. There are such huge numbers of various sorts of algorithms in ML which are mainly Supervised and Unsupervised Learning. In case of supervised learning, the training data contain both features and labels, where features are the values on which model is trained and labels are the output value. Classification is one of the typical task supervised learning. The spam filter is a genuine case of this. Another difficult task is to predict a target numeric value, for example, the cost of a vehicle, given a set of features (age, brands, mileage, and so on.) called predictors. That kind of job is known as regression. In unsupervised learning, as you might guess because the training data doesn't have labels, also not classified and analysis of data is done without giving any instructions. The model tries to learn without the teacher. In our Breast Cancer dataset, we have two sets of Dependent variables i.e., Malign, denoted as M or Benign, denoted as B. So, in this experiment we use the supervised learning classification algorithm. We have picked three distinct types of ML classification algorithms.

3.5 *Logistic Regression*

It is the regression investigation to direct when the dependent variable Y i.e., output data is categorical or binary. Like all regression analysis is also used to improve the link between one dependent variable and at least one independent variables, it's a predictive analysis and used to explain data. The Binary Logistic Regression has 2 categories and characteristics are at 2 levels. Eg- Pass or Fail, Yes or No, High or Low. If the expected probability is over half, by then it gives the case with the high class (called the positive class, marked "1"), or, other than it gives the low class (i.e., it has a place with the negative class, named "0"). It predicts in the case of something is true '1' or false '0' as opposed to predicting something continuous like the size [4]. Logistic Regression model estimated probability given as-

$$\hat{p} = h(x) = \sigma(\theta T \cdot x) \qquad \text{...(1) [5]}$$

It is called the logit noted, where σ(.) is a sigmoid function that give the output as a number somewhere in the range of 0 and 1. It is represent as shown in the below equation and Figure 1.

$$\sigma(t) = \frac{1}{(1+\exp(-t))} \qquad \text{...(2) [9]}$$

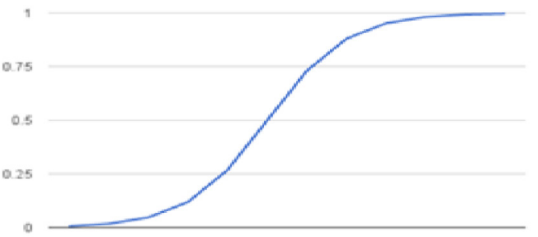

Figure 1

Logistic Function [6]

Plot above is transformed into the range 0 and using logistic function and e is called the base of natural logarithms. It is an S-shape curve, it used to take any actual values and draft it into somewhere in the between 0 and 1.

3.6 *Support Vector Machine*

The SVM model is a very adaptable and well-known model in the world of Supervised Learning algorithms and algorithms in general. Performing linear or non-linear classification, regression, and even outlier

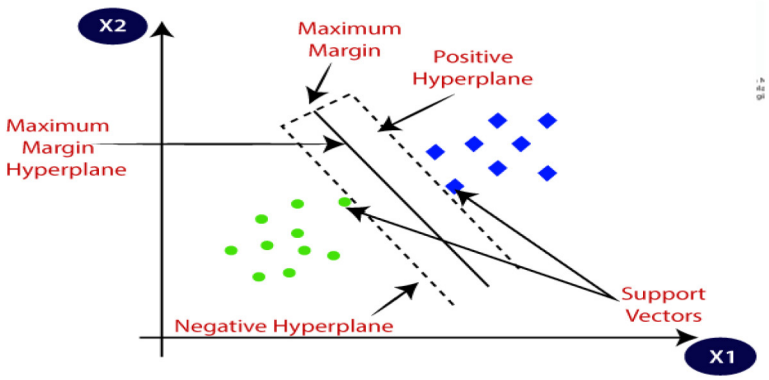

Figure 2

Classification of data through Hyperplane [10]

identification are all possible with this tool. Based on the principle of categorising or categorising the dataset into two categories or classes, this binary classifier uses a straight-line Y=mX+C in n-dimensional space to separate the data [7]. As a result, by simply placing the plane in the plane, the class of new dataset can be quickly determined. This optimal choice boundary is referred to as a Hyperplane, and it is responsible for completing the task of categorising the individuals into groups.

When WTX=0, the Hyper Plane is in the same state as when Y=mX+c, which is the condition of a straight line, where W, X is indicated by a vector. Consider the above figure 2, which depicts two separate classes, each of which is distinguished by the use of a decision limit or a hyperplane:

SVM may be divided into two types: linear SVM and non-linear SVM. Linear SVM is the more common of the two types. [8] In Linear, SVM datasets that are linearly distinguishable by using a straight line are referred to as linear information, and datasets that are not characterised by using a straight line are referred to as non-linear information, and the classifier used is referred to as Non-Linear SVM classifier.

3.7 *Random Forest*

Known as Random Forest, it is one of the most spectacular and well-known machine learning algorithms currently available. It is capable of carrying out both the Classification and the Regression tasks simultaneously. It requires less training time than other algorithms and produces output with excellent accuracy, even when dealing with

enormous datasets. It is also efficient when dealing with small datasets. It may also maintain accuracy even when a large amount of information is not available at the time. It depends on the idea of ensemble learning as the name recommends this calculation make the forest with some of the decision trees. When all is said in done, the more tree in the forest the more robust the prediction and in this way the higher accuracy. In the random forest, we develop different trees as help to a solitary tree. To classify another item dependent on an attribute, every tree gives a prediction and we state this tree gives votes in favor of the particular class. The forest picks the class has been the most maximum vote across different trees in the forest also on account of regression it always takes the aggregate of the yield by various trees [9]. We can grasp the working of the Random Forest calculation with the following steps-

- Step 1 – The planning assumes that numerous situations are N. As a result, the sample of these N samples is picked at random and substituted.
- Step 2 – In the event that M input variables are present, a number mM is resolved with the final goal of selecting a random m variable from M input variables at each node. This m split is used to divide the node, and the estimation of m is remained steady while we develop the forest.
- Step 3 – Without trimming, a tree's appearance is at its most impressive.
- Step 4 – Determining new information by conglomerating the predictions of the n tree trees (for example greater part votes in favor of classification, the normal for regression). Figure 3 shows Flow Chart of Random Forest.

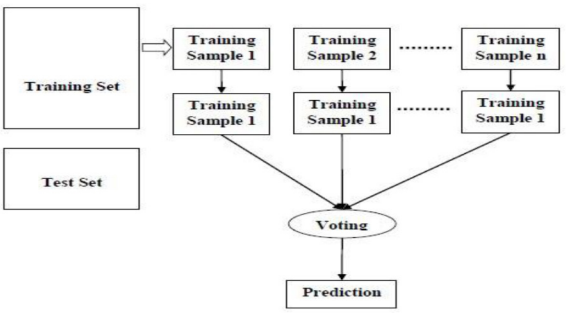

Figure 3

Flow Chat of Random Forest [6]

4. Results and Discussion

UCI Breast cancer dataset consists of 32 attributes, so dimensionality reduction plays an important role in reducing the multidimensional data to a couple of dimensions. We applied each of the three supervised algorithms Random Forest Classifier, Support Vector Machine, Logistic Regression, and the result shows that the highest accuracy of 98.6% is achieved by Random Forest Classifier when a comparative study was done with other algorithms. That's why we recommend that Random Forest is the most appropriate algorithm for the classification of Breast Cancer events with complex datasets. To represent the accuracy, we generally use the confusion matrix and it's calculated as-

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \quad ...(3)\ [7]$$

Table 1

Comparison of various algorithms

Algorithm	Accuracy	Precision	Recall
Random Forest	98.60	0.99	0.98
Logistic Regression	94.41	0.94	0.94
Support Vector Machine	93.71	0.93	0.93

Table 1 Shows the Comparison of various algorithms.

Conclusion

Now a day's many methods like Machine Learning and Data Mining are utilized to analyze the medical field data. The main challenge with these methods is to construct accuracy and computationally proficient classifiers for clinical applications. Our research, for the most part, centered around the improvement of predictive models to accomplish great accuracy in foreseeing substantial infection utilizing administered learning strategies. Future research right now is done for upgrading the performance of the classification technique with the goal that it can predict on more factors.

References

[1] Ch. Shravya, K. Pravalika, ShaikSubhani," Prediction of Breast Cancer Using Supervised Machine Learning Techniques", *International*

Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6 (April 2019).

[2] Hiba Asria*,Hajar Mousannifb, Hassan Al Moatassimec, Thomas Noeld," Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science, Vol- 83, PP-1064 –1069 (Year-2016).

[3] Haifeng Wang and Sang Won Yoon – "Breast Cancer Prediction using Data Mining Method", IEEE Conference Proceedings of the 2015 Industrial and Systems Engineering Research Conference (2015).

[4] B.M.Gayathri, C.P.Sumathi, T.Santhanam," Breast Cancer Diagnosis Using Machine Learning Algorithms– A Survey", *International Journal of Distributed and Parallel Systems* (IJDPS) Vol. 4, No. 3 (May 2013).

[5] T.L. Octaviani and Z. Rustam, "Random Forest for Breast Cancer Prediction", Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018) (Year-2018), AIP Conf. Proc. 20168, 020050-1-0200050-6.

[6] "Classification Algorithms- Random Forest", Available at https://www.tutorialspoint.com2/machine_learning_with_python/machine_learning_with_python_cl assification_algorithms_random_forest.htm .

[7] "Breast Cancer Detection Using Machine Learning", Available at-https://medium.com/@randerson112358/breast-cancer-detection-using-machine-learning38820fe98982.

[8] "Building a Simple Machine Learning Model on Breast Cancer Data", Available at-https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-dataeca4b3b99fa3.

[9] Aurelien Geron, "Hands-On Machine learning with Scikit-Learn &TensorFlow", O'REILLY Media (Year-2018).

[10] "Support Vector Machine Algorithm", Available at- https://www.javatpoint.com/machine-learningsupport-vector-machine-algorithm.