International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217141

Detection of Breast Cancer Using Machine Learning Algorithms

Vijaylaxmi Kochari
Bharatesh College of Computer Applications, Belagavi, Karnataka, India

Article Info
Volume 7, Issue 1
Page Number: 223-227
Publication Issue :
January-February-2021
Article History
Accepted : 01 Feb 2021
Published : 08 Feb 2021

ABSTRACT

Breast cancer represents one of the dangerous diseases that causes a high number of deaths every year. The dataset containing the features present in the CSV format is used to identify whether the digitalized image is benign or malignant. The machine learning models such as Linear Regression, Decision Tree, Radom Forest are trained with the training dataset and used to classify. The accuracy of these classifiers is compared to get the best model. This will help the doctors to give proper treatment at the initial stage and save their lives.

Keywords: Breast Cancer, Linear Regression, Decision Tree, Random Forest.

I. INTRODUCTION

There are more than one million breast cancer cases are found every year. This is the main cause of the high rate of death yearly. It is a type of cancer that occurs mostly in females and is the leading cause of women's deaths. The dataset used in this work contains features that are computed from a digitized image of a fine needle aspiration (FNA) biopsy of a breast mass [1]. They describe the characteristics of the cell nuclei present in the image at cancer affected part. The diagnosis of breast cancer is done by classifying the tumor. Tumors can be either benign or malignant. Malignant tumors are more harmful than benign.

The breast cancer symptoms are skin irritation, nipple pain, change in breast color (like red, brown), increases in breast size or shape at a short period of time, swelling in the part of the breast, lump or mass in the breast. The disease symptoms are not presented well in advance and hence diagnosis is delayed. For many years, the X-ray was the only method that was used to detect breast cancer. There is a self-test every woman can do it regularly using her hand to check for any abnormal growing cells, another way is going to a specialist doctor for mammogram tests.

So, to detect the cancer in the initial stage we can make use of Machine Learning. Machine learning algorithms are used to predict the type of cancerous cells effectively and accurately [2]. Machine learning is a subset of AI that provides the system with the ability to automatically learn to improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access the data and learn themselves.

Machine learning is divided into supervised, unsupervised, semi-supervised and reinforcement machine learning. In supervised learning, a set of data

instances is used to train the machine and are labeled to give the correct result. In unsupervised learning, there are no pre-determined data sets and no notion of the expected outcome. In semi-supervised machine learning is a combination of supervised and unsupervised machine learning methods. In Reinforcement Learning the learning process continues from the environment in an iterative fashion. All possible system states are eventually learned by the system over a prolonged period of time [3].

In this paper, we are making use of Supervised machine learning models namely Linear Regression, Decision Tree and Random Forest.

II. SYSTEM DESIGN

A. Data Flow Diagram

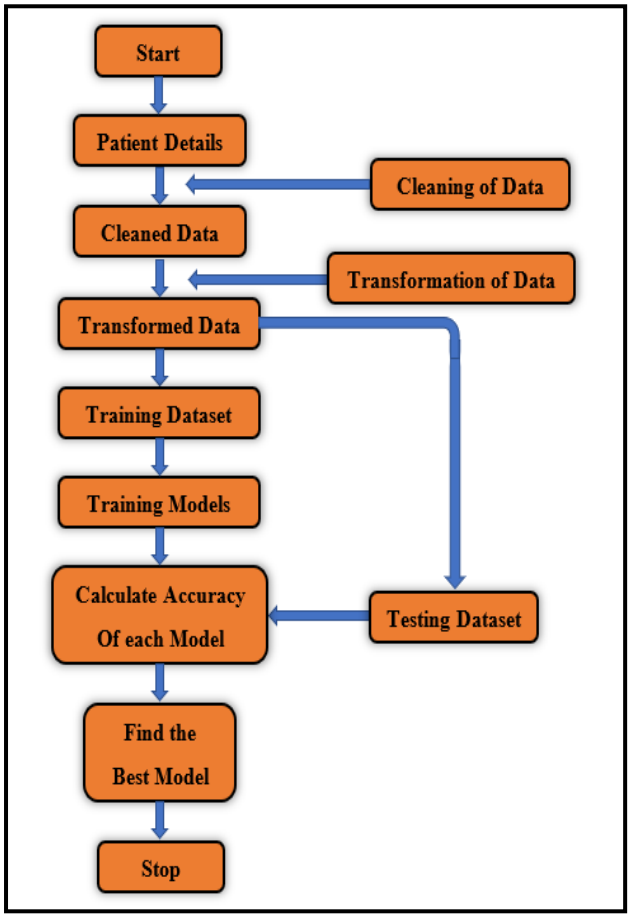

Figure 1 : Data Flow Diagram

B. Dataset

Dataset is a collection of data; the dataset is used to train the model for performing various actions. We have collected the patient data from the Kaggle in the CSV format. In CSV format information is separated by commas and saved with .csv extension.

This dataset has a total 569 Patient details, consists of 33 attributes few of them are ID Number, Diagnosis (M=malignant, B=benign), Radius, Texture, Perimeter, Area, Smoothness, Compactness, etc.

C. Cleaning of Data

The patients' dataset from the Kaggle in the CSV format has few attribute values as NA, Null, NAN, etc. We can remove these unwanted data from the dataset. We can get a count of the number of Malignant (M) or Benign (B) cells using the ['diagnosis'].value counts () method. Out of 569 patients' details, 212 are counted as Malignant and 357 as Benign.

D. Transformation of Data

The next step is the transformation of data because the machine doesn't understand the character or strings. So, we transform the cleaned data into numbers using the sklearn libraries.

E. Spliting of Dataset

After the transformation dataset will be divided into two parts called training and testing dataset. We can split the dataset into a 75% Training dataset and a 25% Testing dataset. We train the machine learning models Linear Regression, Decision Tree and Random Forest using the training dataset. Then we use the testing dataset to compare the accuracy of the machine learning models to find out the best model.

F. Models

1) Linear Regression

Linear regression is based on a supervised machine learning model and it performs a regression task.

Regression models target prediction values based on independent variables. It is used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between the dependent and independent variables, they are considering the number of independent variables being used.
This model performs the task to predict a dependent variable value(y) based on a given independent variable(x). So, this regression technique finds out a linear relationship between x and y. Hence, the name is linear regression.

It is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the dataset we have, with the belief that those outputs would fall on the line.

2) Decision Tree

The decision tree is one of the supervised machine learning model. They can be used to solve classification as well as regression problems. The tree representation is used by it, to solve the problem in which each leaf node corresponds to a class label and attribute are represented on the internal node of the tree. The goal is to create a model that predicts the value of a target variable by learning simple decision rules deduced from the features of data.
We can represent any Boolean function on discrete attributes using a decision tree model and its major challenge is the identification of the attribute for the root node at each level. This process is known as attribute selection.
For approximating discrete-valued target functions, Decision tree learning method is used. It makes use of decision tree to represent the learned function. Decision tree learning is one of the most widely used and practical methods for inductive inference [4].

3) Random Forest

Random forest is based on a supervised machine learning model which is used for both classifications as well as regression. As we know that a forest is made up of trees and more trees mean more robust forests. The random forest creates a decision tree on the data sample and then gets the prediction from each of them and finally selects the best solution.
Random forest classifier will handle the missing values and it maintains the accuracy of a large proportion of data [5]. If there are more trees, it won't allow overfitting trees in the model.
We can understand the working of the random forest model with the help of following steps:
Step1: First, start with the selection of random samples from a given dataset.
Step2: It will construct a decision tree for every sample. Then it will get the prediction result from every decision tree[6].
Step3: It will be performed for every predicated result.
Step4: Last, select the most voted prediction result as the final prediction result.

G. Confusion Matrix

To get an accuracy confusion matrix can be used. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values broken down by each class. This is key to the confusion matrix. The confusion matrix contains notations: TP, TN, FP, and FN [7].
The meaning of these notations are:
True Positive (TP): Observation is positive and it is predicted to be positive.
True Negative (TN): Observation is negative and it is predicted to be negative.
False-Positive (FP): Observation is negative, but it is predicated positive.
False-Negative (FN): Observation is positive, but it is predicted negative.
Syntax of Confusion Matrix:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

H. Classification Rate/Accuracy

Classification rate or accuracy[8] is given by the relation (i):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{--------- (i)}$$

Training of Linear Regression, Decision Tree and Random Forest models are done using 426 patients' dataset, testing is done with 143 patients' dataset.

The Table 1 clearly depicts that **Linear Regression model** correctly classifies 135 patients' dataset as benign or malignant and 8 patients' dataset wrongly as benign or malignant. So, the accuracy of the Linear Regression model is 94.41%.

Confusion Matrix of Linear Regression model:

Confusion Matrix of Linear Regression model:

Predicted Cancer Status

Actual Cancer Status	Total = 143	Benign	Malignant	
	Benign	TN = 86	FP = 4	90
	Malignant	FN = 4	TP = 49	53
		90	53	Total = 143

Table 1 performance analysis using Linear Regression model

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (49 + 86) / (49 + 86 + 4 + 4)
= (135) / (143)
= 0.9441

Linear Regression Testing Accuracy = 94.41%

The Table 2 clearly depicts that **Decision Tree model** correctly classifies 136 patients' dataset as benign or malignant and 7 patients' dataset wrongly as benign or malignant. So, the accuracy of the Decision Tree model is 95.10%.

Confusion Matrix of Decision Tree model:

Predicted Cancer Status

Actual Cancer Status	Total = 143	Benign	Malignant	
	Benign	TN = 84	FP = 6	90
	Malignant	FN = 1	TP = 52	53
		85	58	Total = 143

Table 2 performance analysis using Decision Tree model

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (52 + 84) / (52 + 84 + 6 + 1)
= (136) / (143)
= 0.9510

Decision Tree Testing Accuracy = 95.10%

The Table 3 clearly depicts that **Random Forest model** correctly classifies 138 patients' dataset as benign or malignant and 5 patients' dataset wrongly as benign or malignant. So, the accuracy of the Random Forest model is 96.50%.

Confusion Matrix of Random Forest model:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (51 + 87) / (51 + 87 + 3+ 2)
= (138) / (143)
= 0.9650

Random Forest Testing Accuracy = 96.50%

From the above results it is clear that accuracy of Linear Regression is 94.41%, Decision Tree is 95.10% and Random Forest is 96.50%. From these results we can depict that among these models Random Forest is the best model for the breast cancer detection.

III CONCLUSION

Machine Learning techniques have been widely used in the medical field and have served as a useful diagnostic tool that helps physicians in analysing the available data

as well as designing medical expert systems. This model contains three machine learning techniques Linear regression, Decision tree, Random forest. These models take the cleaned training dataset and learn from that. Later the testing dataset is given to these models to check the accuracy. So, the accuracy of the Linear regression model is 94.41%, Decision Tree is 95.10% and Random Forest is 96.50%. So, from this, it is clear that the accuracy of Random Forest is more compared to Linear Regression and Decision Tree. So Random Forest is the best model for the detection of breast cancer. It will be helpful for doctors to give treatment to patients. Patients can have a better chance to live their life considerably, ensuring a good quality of life.

III. REFERENCES

[1]. A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.

[2]. B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5, doi: 10.1109/ICCIC.2016.7919576.

[3]. S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

[4]. D. Wang and L. Jiang, "An Improved Attribute Selection Measure for Decision Tree Induction," Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), Haikou, 2007, pp. 654-658, doi: 10.1109/FSKD.2007.161.

[5]. Vrushali Y Kulkarni, Pradeep K Sinha, "Efficient Learning of Random Forest Classifier using Disjoint Partitioning Approach" Proceedings of the World Congress on Engineering 2013 Vol II, WCE 2013, July 3 - 5, 2013, London, U.K. ISBN: 978-988-19252-8-2 ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online).

[6]. Shubham Pawar "Prediction of Movie Performance using Machine Learning Algorithms", International Journal for Research in Applied Science and Engineering Technology. 8. 667-672. 10.22214/ijraset.2020.2102.

[7]. P. Varaprasada Rao, S. Govinda Rao, P. Chandrasekhar Reddy, B. S. Anil Kumar, G. Anil Kumar, "Detection of Malicious uniform Resource Locator" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019.

[8]. Vijaylaxmi K. Kochari "Multi-Model Analysis of Mammograms" International Journal of Computer Sciences and Engineering - Vol. 9, Issue.1, January 2021, E-ISSN: 2347-2693.

Cite this article as :

Vijaylaxmi Kochari, "Detection of Breast Cancer Using Machine Learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 223-227, January-February 2021. Available at doi : https://doi.org/10.32628/CSEIT217141
Journal URL : https://ijsrcseit.com/CSEIT217141