TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

Predictive Diagnosis of Cancer using Machine Learning.

R. Anushya, S. Amritha, S. Yuvasakthi, Mrs. Ranjana

Article Info
Volume 83
Page Number:7558 - 7567
Publication Issue:
May-June 2020

***Abstract*:**

In Today's world, Technology has revolutionized almost every aspect. However,Despite the advancements various diseases are still in the rise. Among them , Cancer stands as one of the major cause of death and accounts for about 9.6 million deaths worldwide . However, current evidence suggests that by the introduction of a proactive system which incorporates the avoidance and modification of key risk factors and by introducing a mechanism which enables detecting and predicting cancer at the earliest stage accurately would pull down the cancer death percentage by 30-50%. Technology-enabled smart healthcare is no longer a flight of fancy. However, The required facility for diagnosing cancer accurately and at the earliest stage using the results of biopsy are not available to all general hospitals. Identifying and diagnosing cancer at the earliest stage is crucial as the possibility of cancer spreading increases. Therefore, A computerized system which identifies cancer at the earliest stage with minimal time with greatest accuracy and which reduces cancer recurrence and mortality has to be developed. This paper concentrates and summarises the different machine learning algorithms which may be implied in cancer diagnosis to improve the accuracy of the diagnosis and identification.

Objective: Though numerous data are available in the medical field most of the data are not analysed for capturing valuable knowledges. Advanced techniques could be used to discover patterns and its relations. Our study shows the implementation various of machine learning algorithms for developing a breast cancer predictive model..

Method: The primary objectiveis to implement various ML techniques such as Support Vector Machine (SVM) , Random Forest (RF) and Decision Tree(DT) to develop the predictive model and to compare the parameters such as accuracy and performance of the different algorithms.

Result and Conclusions: The results of our analysis indicate the accuracy of SVM, RF and DT as 0.98, 0.9813 and 0.95 respectively.Based on the our results we concluded that RF classification model predicts with highest accuracy and least error rate in comparison with the other algorithms for breast cancer. It was also found that DT showcased the least acccurate prediction model for breast cancer.

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *18 May 2020*

INTRODUCTION

Cancer is a deadly disease which can occur in various parts of the body. Among them Breast cancer(BC) stands as one of the most common cancer which more commonly occurs for women. Early detection of BC can greatly improve prognosis and survival chances by providing clinical treatment to patients early. Cancer is defined as unregulated growth of cells in the body. The cell which divides

and grows in uncontrollable form resulting in an abnormal mass of tissue called tumour. Tumour can disrupt the normal functioning. However, not every tumour erupting in the body is due to cancer.

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

Classification of cancer based on the affected cell and presently close to 200 different types of cancer are present.Our paper focuses on breast cancer. Breast cancer is common among females across the world [2]. Recent years show the increased improved rate in survival of women from breast cancer due to extensive screening and advanced treatments.

Predicting outcome is a challenging task while data mining technique helps to simplify the prediction segment and automated tools help to make it possible for larger dataset. This technique makes it possible to predict results of a disease from pre existential datasets.

OVERVIEW OF BREAST CANCER

Breast Cancer [1] has its genesis from breast tissue mostly among inner linings of ducts and lobules. It occurs in women commonly and rarely in men. The risk of dying from this disease is higher due to the lack of early detection, even with enhanced treatment. Statistics disclose that there were 40,000 female deaths. In 2014 there were 232,670 new cases recorded in the United States [2].

Types of Breast Cancer

1. Benign Breast Cancer (Non-Invasive) [3]: It is also known as carcinoma in situ. As the name indicates this disease remains entirely in its place of origin (insitu) and doesn't spread to nearby tissue regions. Ductal carcinoma in situ (DCIS) cancer type is grown usually inside the milk duct. DCIS is developed in both men and women.

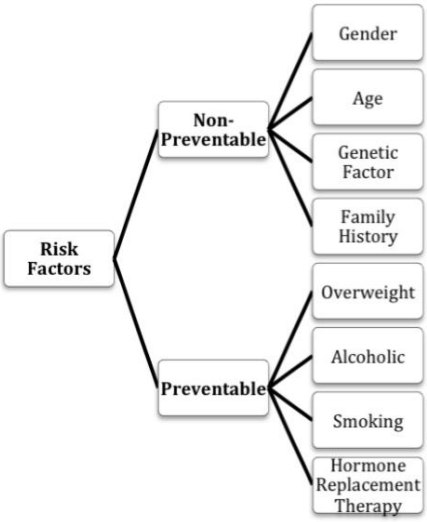
2. Malignant Breast Cancer (Invasive) [3]: This type has an ability to spread to the nearby tissues and is a threat to life. Invasive ductal cancer is one of the most common types of invasive breast cancer. 80% of breast cancer comes under this type.

3. Other types of Breast Cancer [3]: This is the least common type of breast cancer which includes invasive lobular breast cancer(which develops in the cells of milk producing lobules), tubular breast cancer, inflammatory breast cancer, papillary breast cancer and medullary breast cancer.

Risk Factors

Something that affects the individual to acquire some disease, such as breast cancer is known as a risk factor.

There have been many cases where women have breast cancer without apparent risk factors.



It can be classified based on non-preventable and preventable [4]

Non-Preventable Risk Factors:

1. Gender: In this case main risk is being a woman. The disease is likely to occur in women about 100 times more rather than men.

2. Age: Age also plays a vital role in growing cancer. Invasive breast cancers

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

are developed in women with age 55 or older than that.

3. Genetic risk factors: Due to strong genetic risks 5% of cancer are found to be inherited to an individual. There are two autosomal dominant genes, BRCA1 and BRCA2 that account for most cases of familial breast cancer. 65% to 85% of women with harmful BRCA mutation have risk of developing breast cancer.

4. Family History: If the woman's mother, sister, father or child has been diagnosed with breast or ovarian cancer then risk of growing the disease increases(two-fold). Even if the relative was diagnosed before the age of 50 the risk increases.

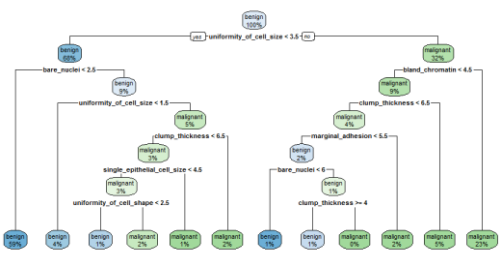
Preventable Risk Factors:

Alcoholic consumption, smoking and being overweight has high risk of recurrence. Other risk factors include High BMI after menopause: Lack of exercise, Radiation Therapy to the chest (before age 30), Hormonal use – postmenopausal ,Weight gain after menopause, Race (African American-higher risk), High bone density, , Late pregnancy at an older age.

METHODOLOGY

Decision Trees

[7] A classifier that is represented as a recursive partition of the instance space is known as decision tree. A predictive model is created that maps observations about a node to conclusions regarding the target value of the nodes.



For breast cancer diagnosis a predictive model is developed using a decision tree. In this paper a decision tree is used for classification to establish the model. [9] To prepare training and test data 10-fold cross validation is used. After data pre-processing (CSV format) using R studio, the algorithm is employed on the dataset, subsequently the data are classified into "benign" or "malignant" based on the endmost outcome of the decision tree that is constructed.

For conducting the procedure algorithm is as follows:

ALGORITHM:

- INPUT: Wisconsin Breast Cancer data set is taken as the input.
- OUTPUT: Decision Tree Predictive Model with leaf node either benign or malignant using classification tree.

PROCEDURE:

1. Acquire dataset.

2.For applying classification tree technique, pre-process the data first.

3.After which the pre-processed dataset is uploaded in R studio for analysis.

4.Implementing the decision tree algorithm, a decision tree with leaf nodes consisting of class tag as benign and malignant is generated.

5. By cross referencing new attribute values in the decision tree, the diagnosis of new patients is attained and the tumour is specified by tracing the path till the leaf node (i.e. benign or malignant).

Data Description and Pre-Processing

From the UCI Machine Learning Repository [10] the Wisconsin Breast

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

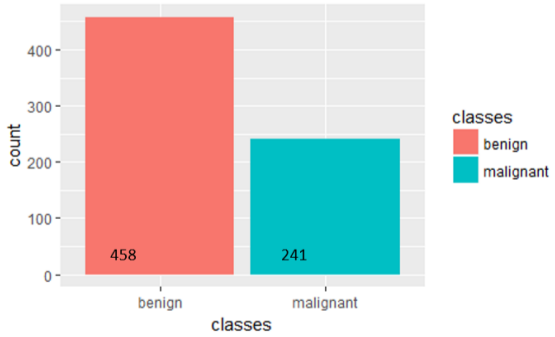
Cancer datasets are taken which is then used to separate non-cancerous from cancerous samples.

Dataset	No. Of Attributes	No. Of Instances	No. Of Classes
Wisconsin Breast Cancer (Original)	11	699	2

Wisconsin Breast Cancer Dataset Attribute

S.No	Attribute	Domain
1	Sample Code Number	Id number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2(Benign) or 4(Malignant)

699 instances of breast cancer patients with each, either having cancerous (malignant) or non-cancerous (benign) type of tumour is contained within the dataset. In the dataset, each record of the last attribute was changed. If the attribute value is equal to 2 it was changed to Benign and to Malignant if the attribute value is equal to 4.



Results and Discussion:

To classify whether a patient had benign or malignant tumour, the tree generated by classification tree algorithm can be used. The information entropy concept is used in the data mining technique. To split the data into smaller modules, each attribute of the data is used.

The rpart package is employed in this study, it implements the classification and to build DT for prediction as well as evaluation of the all data, regression tree (CART) function is used. The regression tree function processed the input, yielded the model accuracy and an optimal tree as the final result. At the top of the tree the DT contains a root node to denote the main variable, along with decision nodes and terminal nodes with percentages of classification. DT is selected as one of the algorithms to evaluate the data as it is known to handle various types of data [11,12,13,14].

Confusion Matrix of Decision Tree

	A - Benign	B-Malignant
A - Benign	438 (a)	20 (b)
B-Malignant	18 (c)	223 (d)

Correctly Classified Instances (661) 94.5637 %

Wrongly Classified Instances (38) 5.4363%

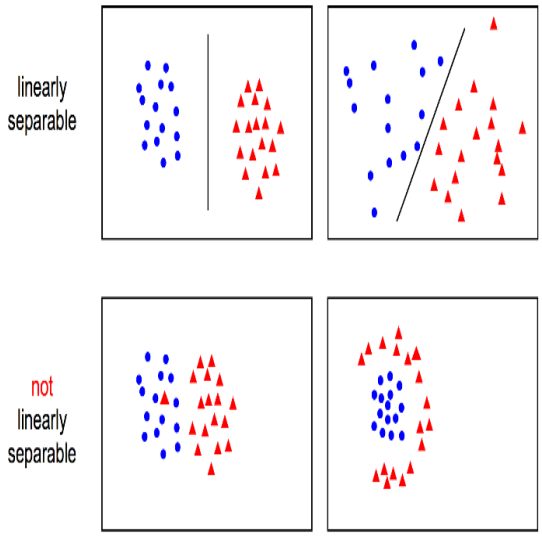
Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a binary linear classification in which decision boundariesare constructed explicitly to reduce generalization error. It is one of the very powerful and versatile Machine Learning models which is capable of performing both the linear or nonlinear classification, regression and even outlier detection [21]. SVM models are particularly well suited for small and medium sized complex objects.

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

SVM classification

SVM is a supervised machine learning algorithm which is completely based on the classification procedure.In this we plot each data item as the points in n-dimensional space with each feature value as the value of particular coordinates.And then the classification is performed by identifying the hyperplane that differentiate the classes. If classification is "linearly separable", SVM fits the "decision boundary" that is defined by the largest margin between the closest points for each class. This is called "maximum margin hyperplane (MMH)".



Model Training
From our dataset, let's create the target and predictor matrix

"y" =feature which we are trying to predict (Output). Here our "target" is to find whether the data are cancerous (Malignant) or not (Benign).

"X" = the remaining column which act as the predictors(mean radius, mean texture, mean perimeter, mean area, mean smoothness, etc.)

Create the training and testing data

Now that we've assigned values to our "X" and "Y", the next step is to import the R library that will help us split our dataset into training and testing data.

Training data = the subset of our data used to train our model.

Testing data = the subset of our data that the model hasn't seen before (We will use this dataset to test the performance of our model).

Splitting our data using 80% for training and the remaining 20% for testing.

The size of training "X" (input feature) :(455,30)

The size of testing "X" (input feature) :(144,30)

The size of training "Y" (output feature) :(455,)

The size of testing "Y" (output feature) :(144,)

First step involves importing the svm model and training it using the training dataset. Then with the training it predicts the output of the test dataset. Next step is to check the accuracy of our prediction by comparing it to the output we already have (y_test), using a confusion matrix for this comparison. And then Improving our model. Then normalization is carried out which is a feature scaling process that brings all values into range [0,1]

Confusion Matrix on Scaled dataset

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

```
Confusion Matrix and Statistics


test_pred_grid benign malignant
     benign       153         4
     malignant      2        80

               Accuracy : 0.9749
                 95% CI : (0.9462, 0.9907)
    No Information Rate : 0.6485
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9446

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.9871
            Specificity : 0.9524
         Pos Pred Value : 0.9745
         Neg Pred Value : 0.9756
             Prevalence : 0.6485
         Detection Rate : 0.6402
   Detection Prevalence : 0.6569
      Balanced Accuracy : 0.9697

       'Positive' Class : benign

> |
```

The accuracy achieved is 98%.

Random Forest

Random forest algorithm is a supervised classification algorithm as the name suggests it creates several classification by several ways and makes it random [21].It builds numbers of Decision trees using random samples with a replacement to overcome the problem of DTs. Each tree classifies its observations, and the majority voted classification decision is chosen. RF is used in the random mode for assessing proximities among data points.

Random forest predictions are based on the generation of multiple classification trees. They can be used for both classification and regression tasks. This paper used classification tasks. We access the cross-validation results with model_rf$pred by specifying savePrediction=TRUE. After predicting the feature importance and estimating the importance the predicted data for the test data's confusion matrix is obtained.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  benign malignant
##   benign       133         2
##   malignant      4        70
##
##                Accuracy : 0.9713
##                  95% CI : (0.9386, 0.9894)
##     No Information Rate : 0.6555
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9369
##  Mcnemar's Test P-Value : 0.6831
##
##             Sensitivity : 0.9708
##             Specificity : 0.9722
##          Pos Pred Value : 0.9852
##          Neg Pred Value : 0.9459
##              Prevalence : 0.6555
##          Detection Rate : 0.6364
##    Detection Prevalence : 0.6459
##       Balanced Accuracy : 0.9715
##
##        'Positive' Class : benign
##
```

Random Forest Accuracy

Accuracy:98%.

Previous Related Work: -

Several studies have been done in machine learning on the basis of the medical field.One in the largest scale is found to be a prediction and diagnosis mechanism using ML[8].Several studies been conducted on cancer prediction and diagnosis using different methods and combined algorithm to increase the accuracy by various features.Most of the studies used SVM to evaluate performance .Their finding shows variance of 95%, range 94%,compactness 86%.Based on their result SVM can be considered to be an appropriate method for cancer prediction.

LITERATURE SURVEY: -

S.NO	NAME OF PAPER	AUTHOR	DATE OF ISSUE	INFERENCE
1.	Application of Machine Learning in Cancer Prediction and	Joseph A.Cruz David S. Wishart	Within the year of 2006	From this paper it absolutely was over that the accuracy of predicting cancer

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

	Prognosis			susceptibleness and mortality is redoubled using machine learning techniques.
2.	Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction	Harry B. Burke, David B. Rosen, Donald Henson	January fourteenth of the year 1995	From this paper it absolutely was over that accuracy of artificial neural networks was beyond TNM staging systems. additionally new prognostic factors may be other to.
3.	Predictive Modelling for the Presence of Prostate Carcinoma using Clinical, Laboratory, and Ultrasound Parameters in Patients with Prostate Specific Antigen Levels < 10 ng/mL	Mark Garzotto, R. Guy Hudson,	2003	From this paper it absolutely was over that increasing the accuracy incorporation of clinical associate degreed TRUS knowledge into a pre diagnostic test representation caused an improvement within the prediction rate of prostate cancer.
4.	The Standardized Uptake Value for F-18 Fluorodeoxyglucose Is a Sensitive Predictive Biomarker for Cervical Cancer Treatment Response and Survival	Elizabeth , Barry A. Siegel.	2007	From this paper it absolutely was over that the SUVmax of cervical tumour was found to be a sensitive biomarker of response to treatment for cervical cancer patients.
5.	Classification and Prediction of Survival in Hepatocellular Carcinoma by Gene Expression Profiling	Ju-Seog Lee	September 2004	From this paper it absolutely was over that the biological variations by organic phenomenon Profiling identified within the HCC subclasses square

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

				measure associate degree supply of development for therapeutic targets.
6.	New blood-based biomarkers for the diagnosis, staging and prognosis of prostate cancer	Shahrukh F	10 August 2007	From this paper it absolutely was over that the long run prognosis of cancer may rely on markers which offer correct molecular staging and indicate the presence of cancer and its stage.
7.	Identification of a predictive gene expression signature of cervical lymph node metastasis in oral squamous cell carcinoma	Su Tien Nguyen	March 19, 2007	In this paper, they need to deal upon the assessment of the cervical lymphatic tissue metastasis standing in rima cancer helps to predict the prognosis of patients and additionally helps to work out the acceptable treatment.
8.	Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models	D.TIMMERMAN	19 July 2010	This paper over that these models facilitate to differentiate between benign and malignant plenty and scale back the chance of wrong prediction.
9.	Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies.	Iliyan Mihaylov	March 2019	Python-based workflow has been developed and therefore the plans for its more improvement are mentioned within the paper.
10.	Prediction of clinical behaviour and treatment for cancers.	Matthias E Futschik	2003	This study demonstrates established clinical parameters with incorporation of microarray knowledge will significantly improve sickness prediction.

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

CONCLUSION:

In our paper, cancer at the side of ML was introduced and studied in addition. an associate degree in-depth literature survey was performed on varied ML ways used for cancer detection. The findings of those researchers recommend that SVM is the most recommended technique used for cancer detection applications. SVM was used either alone or combined with another technique to improve the performance.It was found that SVM has the very best accuracy of 99.8% which will be redoubled to one hundred. Therefore, developing a computerised cancer diagnosis can facilitate to scale back the quantity of your time to diagnose the cancer at the earliest stage with the best accuracy and reduce cancer repetition and mortality. This paper summarizes the survey on varied machine learning algorithms and ways that square measure used to boost the accuracy of predicting cancer at the earliest stage.

REFERENCES:

1. National Cancer Institute: http://www.cancer.gov/cancertopics/type s/breast.
2. National Cancer Institute Breast Cancer, http://www.cancer.gov/cancertopics/type s/breast
3. carcinoma Organization, http://www.breastcancer.org/ symptoms/types
4. carcinoma Organization, http://www.breastcancer.org/ risk/factors/
5. carcinoma Organization, http://www.breastcancer.org/symptoms/
6. Bellaachia Abdelghani and Erhan Guven, “Predicting carcinoma Survivability victimisation data processing Techniques”, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth Siam International Conference on data processing, 06.
7. J. Han and M. Kamber, "Data Mining ideas and Techniques”, Morgan Kaufmann Publishers, 2000.
8. Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, call Tree Analysis on J48 formula for knowledge Mining. Proceedings of International Journal of Advanced analysis in computing and package Engineering, June 2013.
9. H. Blockeel and J. Struyf. economical algorithms for call tree cross-validation. Proceedings of the Eighteenth International Conference on Machine Learning (C. Brodley and A. Danyluk, eds.), Morgan Kaufmann, 2001.
10. William H Wolberg, Olvi Mangasarian, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].
11. Sudhamathi G, Thilagu M, Padmavathi G. Comparative analysis of R package classifiers victimisation carcinoma dataset. Int J Eng Technol. 2016.
12. Chen W. A comparative study of supply model tree, random forest, and classification and regression tree models for spacial prediction of landslide susceptibleness. Catena.
13. Muchlinski D, Siroky D, Kocher M. examination random Forest with supply regression for predicting class-imbalanced war onset knowledge. Polit Anal. 2016.
14. Dong Y, Du B, Zhang L, Member S. Target detection supported random Forest metric learning. IEEE J Sel prime Appl Earth Obs Remote Sens. 2015.

May – June 2020
ISSN: 0193-4120 Page No. 7558 - 7567

15. Golub TR. Molecular categoryification of cancer class discovery and sophistication prediction by factor comparison observation. Science. 1999.
16. TibshiraniR, Hastie T, Narasimhan B, Chu G: diagnosing of multiple cancer varieties by shrunken centroids of organic phenomenon. Proc Natl Acad Sci USA. 2002.
17. Selaru FM. Artificial neural networks distinguish among subtypes of growth body part lesions. Gastroenterol. 2002
18. Liu B, A combinative feature choice and ensemble neural network technique for classification of factor expression knowledge. BMC Bioinf. 2004
19. Alon U, Barkai N, Notterman district attorney, Gissh K, Ybarra S, Mack D, Levine AJ: Broad pattern of organic phenomenon unconcealed by bunch analysis of tumour and traditional colon tissue provided by oligonucleotide arrays. Proc Natl Acad Sci USA. 1999.