iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** III **Month of publication:** March 2019

DOI: http://doi.org/10.22214/ijraset.2019.3142

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

Decision Support System for Breast Cancer Prediction

G. Dhanalakshmi[1], P. Keerthana[2], M. Rohini[3], Yogalakshmi Karunamoorthy[4]
[1]*Associate Professor, Information Technology, Panimalar Institute of Technology*
[2, 3, 4]*UG Scholar, Information Technology, Panimalar Institute Of Technology,*

Abstract: ***Cancer is considered to be one of the deadliest disease in the world. The second largest cancer disease leading to the death in women is Breast Cancer. Breast cancer is one such cancer which develops from the tissues of breast. Early diagnosis of breast cancer is very much necessary in order to prevent the growth of cancer cells. These cancer cells are detected and analyzed using various ML Techniques. In this paper, we have taken the data from Wisconsin Breast Cancer dataset. Five ML algorithms have been applied and a comparison is made between them. The best of five is identified and the accuracy level for the recurrence of cancer in affected patients is also compared. Different types of ML techniques which includes Support Vector Machines (SVMs) and Logistic Regression, Random Forest have been widely applied for research purposes in development of accuracy and better precision. In addition to this, we have also applied the KNN algorithm and Neural Network algorithm to produce better accuracy in analyzing the data.***

Keywords: Supervised Machine Learning; Breast Cancer; Data preprocessing; Feature selection; Cross validation; Classification; Decision Boundary.

I. INTRODUCTION

Breast cancer is considered to be one of the global disease in the world. According to a survey taken in India, 25 yrs. back, in every 100 breast cancer patients, 2% were in 20 to 30 yrs. age group whereas now it has been increased to 4%. Similarly, 7% were in 30 to 40 yrs. back then, now 16%, 22% were in 40 to 50 yrs. previously, 28% now. However, people of the age group 50 to 60 and 60+ have a positive response. 36% of women of age group 50 to 60 were affected by breast cancer which is now reduced to 30%, similarly 33% women of age group 60+ were affected before, which is now reduced to 22%. Increase in number of patients are in the 25 to 40 yrs. of age, and this is very unfortunate disturbing trend. However, many researches have taken step in predicting and preventing cancer, the accuracy part still remains a challenging one. Supportive tools are added to help the physicians to facilitate accurate diagnosis. These tools focus in eliminating the possible diagnostic errors and provide an easiest way for analyzing large amount of data. Machine learning(ML) is the scientific approach in study of algorithms and is also a subset of artificial intelligence which allows the software to predict accurate results by means of experience and without explicitly programmed. Over the last few yrs., ML techniques have been used widely in the development of predictive models to make efficient support system in predicting cancer accuracy. These techniques are used to identify different kinds of patterns in a data set and predict whether a cancer is malignant or benign. The performance of such techniques can be evaluated based on the accuracy of the classification. In this paper, five ML classifiers have been applied to a breast cancer data set and the results are investigated and compared. These techniques are Random Forest (RF), Support Vector Machine (SVM), Logistic Regression, K Neighbors, and Neural Networks(NN). It also focus to validate and evaluate the relationship between the number of features in a data set and performance of the model. The paper is organized as follows: Section II illustrates the different literature survey of previous researches and briefly describes about the algorithms being used. Section III, the description of the dataset is provided. Section IV, methodology used for this study is described. Section V, summary of results and discussions are made to show the accuracy and comparison of the used classifiers. Finally, the conclusion part of this paper is given in Section VI.

II. LITERATURE ANALYSES

A. Related Work

Machine learning techniques have been used widely in the field of medical application for a very long time, mainly in the diagnosis of breast cancer. Many researches have been conducted using similar approach as we are using in our paper. The dataset used in all related researches is collected from the UCI Machine Learning [1] repository as well. One of them includes the research by [2] which used classifiers such as Naïve Bayes with an accuracy of 95.99% and a multilayer Perceptron with an accuracy of 95.29% on the same data sets. Another research conducted by [3] summarizes the applications of machine learning in the field of breast cancer prediction. It also encapsulates the basics of Artificial neural networks, establishes the simplifications required in implementing a

neural network approach and lays out the research work carried out by various other researches in the same. Consequently, it evaluates various models such as Relevance Vector Machines (RVM), Support Vector Machines (SVM) and then compares them on the basis of performance and other metrics such as specificity and sensitivity. The most cogent research was pursued by H. Yusuf [4], in which diagnosis of breast cancer from mammograms was complemented using logistic regression. The data set was not from the UCI Machine Learning repository [2], but instead collected from a survey of questions completed by a radiologist during his observations with cancer patients. From a sample of 130 patients the mammogram result accuracy was 91.5% while accuracy same compared to 46 test samples of validation test, the accuracy obtained was 67.4%. The author reached the conclusion that presence of mass calcifications, skin thickening, and distortions as a result of mammography had high odds of cancer being malignant.

Other researches have also made [5] comparison study of algorithm for predicting the recurrence of breast cancer with the most commonly used algorithms like SVM, RF and Bayesian Networks. There are several other researches that have been conducted in analyzing the severity of the breast cancer. In this paper, we are taking all the conflicts of using a neural network into consideration and we are going to show the accuracy level for predicting the recurrence of cancer in affected people.

B. Support Vector Machine Classifier

Support Vector Machines are used for classification and regression purposes that performs identifying a hyper plane by generating a linear vector. Its main working function is the map the distance between the input vector to a high dimensional plane. The type of SVM used here is linear SVM. The feature set is given as follows, xi, and a label, yi ϟ {0,1}, the minimization of the loss vector is calculated by the support vector machine using the equation below:

$$fi(\theta) = max\ (1 - \theta^{T}x,\ 0)$$

In contrast to other classifiers, there is an additional intercept term that is added to entries by adding the number 1 using the below formula:

$$\min_{\theta} \lambda \ \sum_{n}^{1} fi(\theta) + \|\theta\|$$

Where λ is the penalty parameter that tells the SVM optimization to avoid misclassifying each training example in the data set. For a larger value of this parameter, the optimizer would choose a smaller margin for the hyperplane for correct classification whereas a smaller value of the same would lead to choosing a larger margin for the hyperplane, inadvertently leading to misclassification.

C. Logistic Regression

Logistic Regression is the appropriate regression analysis, it is also a predictive analysis technique in which the output label's variable depends on dichotomous, which means binary. Here Benign or Malignant (B or M) is predicted to be the output label. It indicates the exact position between the dependent and in dependent variable. The graph represent the input variable combined linearly to predict the output value. The logistic function is also called the Sigmoid function and is calculated by using the cost function below [8]:

$$fi^{(\theta)} = p(y_i = 1/X) \ = \ \frac{1}{1+exp(-\theta^{T} x)}$$

D. K Nearest Neighbor Algorithm

K nearest algorithm is a non parametric methods used for classification and regression.it is a type of instance based learning. The best choice of k depends on the data. For each test data point, we would look at the K nearest training data point and take the most frequently occurring classes and assign that class to the test data. Therefore, K represents the number of training data points lying in proximity to the test data point test data The class label for x* is then predicted to be: y^{*} = max

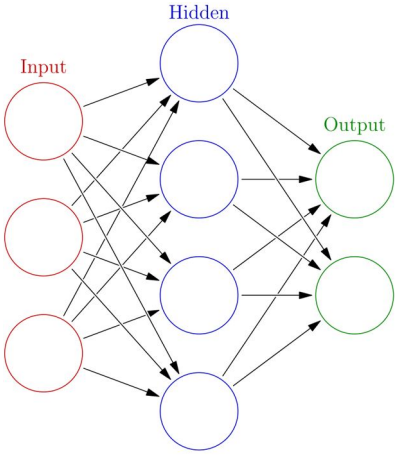
where the ith point has both a vector of features xi and class label yi. For a new point x*, the nearest neighbor classifier first finds the set of neighbors of x*, denoted N(x*) where the indicator function I() is 1 if the argument is true, and 0 otherwise.

E. Random Forest

Random forests are the ensemble learning method used for classification and regression. It is performed by constructing a multitude of decision tree. Random forest correct the habit of over fitting to their training set. random forest classifier creates a set of decision trees from randomly selected subset of training set. Here the trees with high error rate are given low weight value ad vise versa. This would increase the decision impact of trees with low error rate.

F. Neural Network

Neural networks are similar to the biological neural networks that constitute animal brains. it is an paradigm that process the information like the human brain does. the each neurons in the data is analyzed and they are connected in a constrain manner which plays an important role to increase the accuracy in the system.



III. EXPERIMENTAL DATA

The data sets used in this research were procured from the open-source Machine Learning repository of University of California, Irvine [2]. The name of the data set used for the research conducted is Wisconsin Breast Cancer dataset. The dataset was created by Dr. William. H. Wolberg out of the desire to diagnose breast cancer as benign or malignant solely based on the observations of FNA [11]. Further, in collaboration with Prof. Managasarian and two of his graduate students, namely Rudy Sentiono and Kristin Bennett, a classifier was generated [13] that used multi-surface method of pattern separation one the 32 features to finally diagnose 92% of new cases, ultimately producing the Wisconsin Breast cancer data set [14]. the 'class' attribute having a discrete value of either 0 (benign) or 1 (malignant), and the sample code number being a unique patient identification number. The summary of the data sets is provided below:

Table i. Wisconsin diagnosic breast cancer dataset description

S.no.	Attribute/Feature	Range
1.	ID Number	Identification number for patients
2.	Diagnosis	0:Benign, 1:Malignant
3.	Radius	11-27
4.	Area	360-2300
5.	Perimeter	71-82
6.	Texture	11-40
7.	Smoothness	0.05-0.2
8.	Compactness	0.04-.45
9.	Concavity	0.02-0.5
10.	Concave Points	0.02
11.	Symmetry	0.1-0.3
12.	Fractal Dimension	0.05-0.1

Similarly, we preprocess the diagnostic data set in such a way that it can be used to map the patients with recurring cancer. For this purpose, we only extract the patients treated with malignant cancer followed by addition of new attributes. The varies features helps to identify the cancer stage in recurring patients for better understanding and results.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

IV. METHODOLOGY

Diagnosis of breast cancer is a difficult task. The first step involves the clinical examination to detect the tumor/lump in the breast by the General Surgeon or by using imaging techniques such as Mammography which is followed by FNA on the lump detected. The FNA procedure provides with attributes such as tumor size, radius, area etc. These can be used as features for modeling the data into a classifier. Finally, models such as SVM, k-nearest neighbor, Random Forest, Neural network and Logistic Regression are trained and neural network is used to obtain data to make predictions in accurate manner. The general methodology of modeling after data is obtained can be divided into four general steps which can be shown in the figure below. No matter what machine learning algorithm is being implemented, the below mentioned steps are key to any algorithm.

A. Data Preprocessing

This is the first step before modeling the data. At times data obtained may be incomplete, like absence of values in certain rows. It can also be noisy, the data could be replete with errors and outliers, and inconsistent as well, with arrant discrepancies. Thus, in order to eliminate all of the above, we preprocess the data before modeling. Here, we use min max normalization, also known as feature scaling, since the entries inside our data are primarily numeric. It can be described by the following formula [16]:

$$\frac{x\text{-}min(x)}{max(x)\text{-}min(x)}$$

where x is the initial value and min and max are the highest and lowest value respectively.

B. Feature Selection

The selection of right model is completely instinctive. The process depends strongly on the type of data, and the primary aim of the author. Feature selection also known as variable selection, attribute selection or variable subset selection. It is the process of selecting a subset of relevant features(variables, predictors) for use in model construction.it reduces the overfitting,improves accuracy, reduces training time.

C. Cross Validation

At sometimes cross validation is also called as rotation estimation .it is a model evaluation method that is better than residuals. The goal is to predict ad wants to estimate how the accurate predictive model perform. One round of the cross validation involves partitioning a sample of data and involves the training and testing.: the training and test set, split is done in the ratio of 70/30, which means 70% of data is used for training and 30% of data is used for testing. Thus, we select here four models as discussed earlier, and divide both data sets into training and testing sets.

D. Classification

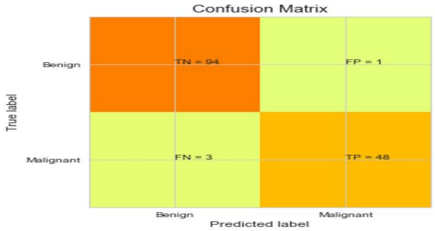
It is a data analysis task that involves the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set categories(sub population), a new observation belongs to ,on the basis of training set of data contain and whose categories membership is known.

Although there are other metrics for model evaluation available as well, classification accuracy is the one most pertinent to our research.

V. RESULTS AND DISCUSSION

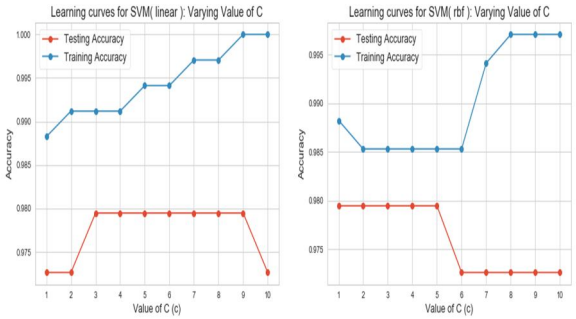
Throughout the research, we closely examine the sensitivity, accuracy, precision, recall using the five different algorithms aforementioned. Form which better accuracy algorithm could be found and the recurrence is identified. The confusion matrix for each model is formulated in order to evaluate the models with ease.

For the Wisconsin breast cancer diagnostic data set, the results were as follows:

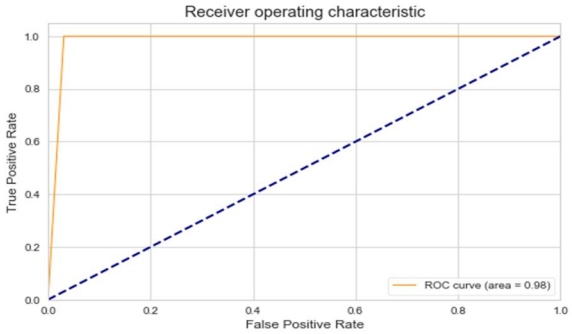


The performance of the classifiers was evaluated using metrics such as accuracy, sensitivity, specificity, and area under curve (AUC) etc. Sensitivity which is also defined as True Positive Rate (TPR) is the percentage of benign tumors data classified as benign by the classifier. The classifier that can correctly classify benign tumors will have a higher result in sensitivity. Sensitivity is defined as follows [15, 17]:

Sensitivity(%)= TPR= $\frac{TP}{TN + FP}X100$



A. Training and testing Accuracy in SVM



Specificity is the percentage of malignant tumors data classified as malignant by the classifiers. The classifier that can correctly classify malignant tumors will have a better result in specificity. It is calculated as follows [18, 19]:

$$\text{Specificity (\%)} = \text{TNR} = \frac{TN}{TN + FP}X100$$

Accuracy is the most crucial metric for model evaluation. It invariably combines specificity and sensitivity for whole of the data combined. It is given by [9,11,13,14]:

$$\text{Accuracy (\%)} = \frac{TP + TN}{TN + FP + TP + FN}X100$$

Area Under Curve (AUC) is a disposition of sensitivity and specificity over all possible thresholds. The AUC value of 100% represents perfect discrimination (the classifier can classify the tumors correctly), whereas an AUC value of 50%.

We see that keeping accuracy as the sole criterion for evaluation, logistic regression is the best classifier model for diagnostic data set . The KNN algorithm gives the accuracy of about 87% .where the neural network is more efficient and provides the accuracy of about 92%.

While we see that the logistic regression classifier performs better for the particular data set, there are some differences in the models which should be kept in mind:

1) Decision boundary: Logistic regression learns a linear classifier, while nearest neighbors can learn non-linear boundaries as well.
2) Predicted values: Logistic regression predicts probabilities, which are a measure of the confidence of prediction. nearest neighbors predict just the labels.

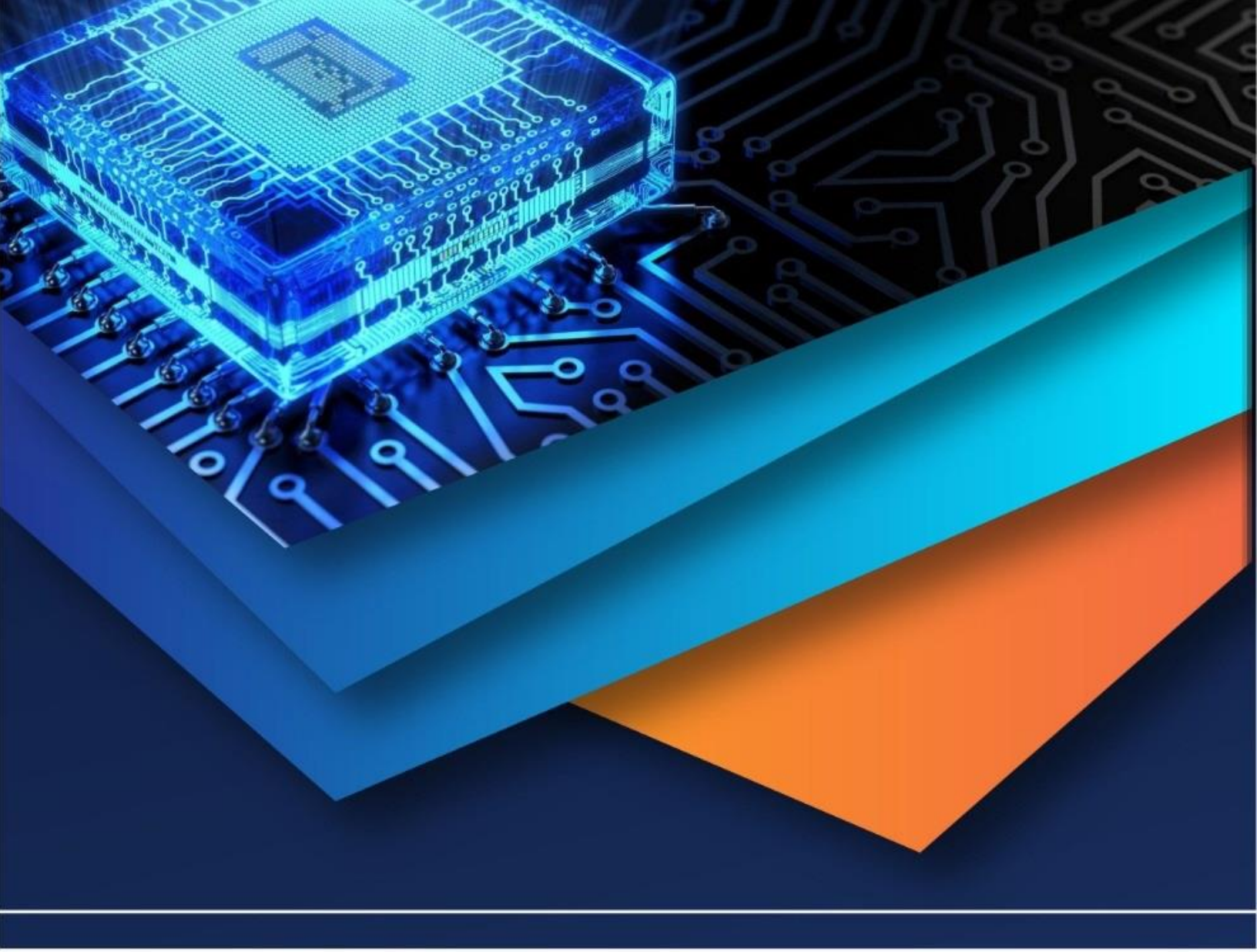
VI. CONCLUSION

While previous researches were not able to achieve an accuracy level with neural network classifier on the same dataset of Wisconsin Breast Cancer, the study proposed here has achieved training accuracies ranging from 93-97% using neural networks. For any model, accuracy cannot be considered as the only metric to be taken into account. The following metrics must be evaluated in order to do a complete evaluation of the model; specificity, sensitivity, area under curve and model accuracy. Furthermore, there may be other metrics as well such as F-score and ROC, which in turn scrutinize the evaluation process.

By using this kind of prototype, we can extend these techniques to hospitals, and maintain a repository of patients with their current diagnosis, and every other detail, which helps in accounting for new cases and used to create the awareness among the patients using machine learning techniques. The use of machine learning techniques cannot be precluded, even the diagnosis rate is made constrain. Use of this prototype should be evaluated seriously before making it use for commercial purpose.

REFERENCES

[1] UCI Machine Learning Repository, Wisconsin Breast cancer data set https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Origi nal), accesed on August, 2016.
[2] Gouda I. Salamal, M.B. Abdelhalim, and Magdy Abd-elghany Zeid, Breast Cancer Diagnosis on Three different datasets using Multi-classifiers, International Journal of Computer and In formation Technology(2277-0764) Volume 01-Issue 01, September 2012
[3] H. Yusuff, N. Mohamad, U.K. Ngah & A.S. Yahaya, Breast Cancer analysis using Logistic Regression, IJRRAS 10 (1), January 2012.
[4] I.S. Jacobs David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 128.
[5] The 2016 IEEE 5th International Conference on Electronic Devices, Systems, and Applications (ICEDSA'2016), December 2016
[6] Mohammed H.Tafisf; Alaa M.El-Hales, 2018 International Conference on Promising Electronic Technologies(ICPET)
[7] Diseases Prediction by Machine Learning Over Big Data From Helthcare Communities, IEEE Access(Volume:5), 26 April 2017.
[8] A study on prediction of breast cancer recurrence using data mining techniques, 2017 7th International Conference on cloud computing, data science & engineering-Confluence.
[9] Machine learning approaches for breast cancer diagnosis and prognosis, 2017 International Conference on Soft Computing and its Engineering Applications(icSoftComp)
[10] Analysis and prediction of breast cancer datasets using data mining classification techniques, 2017 International conference on Intelligent Sustainable Systems(ICISS)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY