IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Comparative Study of Classification Techniques for Breast Cancer Diagnosis

Ajay Kumar[1*], R. Sushil[2], A. K. Tiwari[3]

[1]Dept. of CSE, DIT University, Dehradun, India
[2]Dept. of IT, DIT University, Dehradun, India
[3]Dept. of CSE, KNIT Sultanpur, U.P., India

**Corresponding Author*: *kumarajay7th@gmail.com* *Tel: +91-9557-562-400*

Available online at: www.ijcseonline.org

Accepted: 20/Jan/2019, Published: 31/Jan/2019

Abstract – Classification techniques in Machine Learning are implemented on datasets. In this work, the cancer datasets are used for the classification purpose and collected from UCI Machine Learning repository. There are two types of datasets of breast cancer. Both the datasets are varying by their number of features available across the datasets. This paper presents the implementation and comparative study of major and popular classification techniques such as Decision Tree, k-Nearest Neighbour, Support Vector Machine, Bayesian Network and Naïve Bayes under WEKA environment for accuracy based on evaluation of performance metrics. This paper evaluates that the Bayesian Network gives the best accuracy with less featured dataset while Support Vector Machine gives best accuracy for more featured dataset.

Keywords- Classification Techniques, Feature Selection, k-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Bayesian Network (BN), WEKA

I. INTRODUCTION

As the era is going on with the problem of huge possibilities of carrying cancer because of many reasons. Lots of cancer data is available for research purpose at Kaggle web site, SEER data sheet, UCI etc. Data are available in various formats such as text data, image data, micro array data, gene expression data etc.

Cancer data can be classified into two types: Malignant (M) and Benign (B). It is said that Malignant tumour is dangerous and cancerous when starts growing inside the human body while Benign is less harmful as its cells don't multiply.

The dataset used in this paper is a breast cancer data retrieved from UCI Machine Learning repository and it is a large dataset where data needs to be cleansed for proper utilization. Machine Learning (ML) techniques are used to pre-process the data followed by data cleaning, data selection, finding variable dependency and removal of independent variable. Using ML techniques, breast cancer dataset is classified and labelled for class malignant and benign.
This paper has been organized in six different sections. In the first section, introduction part is detailed. Second part explains about the work done by reputed authors in this domain. In third section, the ML techniques have been discussed for the classification of the data. Fourth section is about the methodologies used for the process of working, in order to get the crisp and error free data for calculation of finding accuracy. Fifth section presents experiment, performed for data analysis. Result have been shown through line chart graph. Sixth section, which is the last part of the paper is followed by references, presents conclusion & future work.

II. RELATED WORKS

Lots of work have been done in the field disease prediction. This section explains about the study of few important research papers.
B. Nithya et al. [1] implemented the three classification methods such as Decision Tree, k-Nearest Neighbour and Naïve Bayes for different datasets viz cancer and Iris datasets in open source R tool environment. The authors also analysed the evaluation metrics like accuracy and error rate. The implementation was focused on type of attributes of dataset and their characteristics.

D. Lavanya et al. [2] used a classification methods Decision Tree to classify the medical data for diagnosis. The author used CART (Classification and Regression Tree) classifier with and without Feature Selection to find the accuracy, time taken to build model for the breast cancer dataset.

Deepika Verma et al. [3] implemented Data Mining technique in WEKA environment to predict the accuracy of diseases like stroke, diabetes, cancer, hypothyroid, heart disease etc. the authors obtained the dataset from UCI Machine Learning repository and applied classification algorithm. The outcomes are obtained 72.7% accuracy on breast cancer dataset and 76.8% accuracy on diabetes datasets.

Morteza H. et al. [5] taken the mammographic image features and build an optimal stratification model for breast cancer risk. The data is taken from 500 women and divided into different age matched class of 50% of higher risk and 50% low risk cases. The feature of images was 44 related to mammographic tissue density distribution between left and right breast. The author built a model LPP (Locally Preserving Projection) based on multi-feature fusion of Machine Learning classification to predict the risk of cancer detection and found 9.7% increase in risk prediction accuracy.

Mahua Nandy [7] implemented supervised and unsupervised learning both in the predicting accuracy of breast cancer. ANN and SVM were from supervised learning where k-Means were from unsupervised learning. SVM outperformed as a result and ANN has disadvantage that it took more time to build the model. The datasets were WBCD and breast cancer dataset with electrical impedance measurement. The author concluded about the feature selection matters a lot to recognise the accuracy.

Md. Milon Islam et al [9] processed the two supervised learning classifiers viz SVM and KNN. They predicted in terms of accuracy, sensitivity, specificity, false discovery rate, false omission rate and Mathews correlation coefficient. They proposed a system with 10-fold cross validation and achieved the accuracy of 98.57% by SVM and 97.14% by KNN on the dataset Wisconsin Breast Cancer Diagnosis.

III. MACHINE LEARNING (ML) TECHNIQUES FOR CLASSIFICATION

There are many algorithms in Machine Learning viz. Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), K-Nearest Neighbours (KNN), Bayesian Network (BN), Decision Tree (DT), and Naïve Bayes (NB) used for data classification. Following we present the key features of each.

A. Support Vector Machine

Support Vector Machine [10,13] is a binary classifier and can be used in classification and regression both. One of the powerful tools when kernel idea is ensembled. It can be applied to predict cancer, database marketing, recommendation system, text categorization, face recognition etc.

B. Multilayer Perceptron

Multilayer Perceptron (MLP) is a classifier based on a Feedforward Artificial Neural Network (FANN) [19]. FANN is a current implementation of spark MLAPI. It is useful when the dataset is small in size. It performs fast operation and easy to implement. It has the features of learning the pattern and generalize the dataset for further operation [17,18]. E.g. Handwritten character recognition.

C. Random Forest

Random Forest [10, 15] is an ensemble classifier based on Decision Tree classifier. RF runs often on large datasets and it is slow in operation compare to other classifiers. RF gives number of classification tree without pruning. The pruning is a technique associated with Classification And Regression Tree (CART) which reduces the size of the tree to find the best predictor in repeating iteration by splitting the data in two subsets. It estimates the missing values and handles the large dataset even with missing values.

D. K-Nearest Neighbours

k-Nearest Neighbour [10,14] is a low costing classifier that uses distance metric. It is also known as Lazy Learning or Instance based learning. The operation is completely done locally with the nearest instance (sample). The distance is measured by Euclidean distance or Manhattan distance method. In this algorithm, the classification is achieved by the minimum distance measured. It does not involve much cost in learning the model but it depends upon the no. of instances where cost increases when no. of instances increase.

E. Bayesian Network

Bayesian Network [20,21] is a directed acyclic graph which represents the probabilistic relationship among variables of interest. Each node represents random (stochastic) variable which has two or more possible state. It infers the probabilistic outcomes numerically from the set of variables on others. It is also known for Belief Network (or Causal Probabilistic Network). E.g. in case of breast cancer, in figure 1, Breast Cancer is represented by the variables and there are two states viz. "present" or "absent".

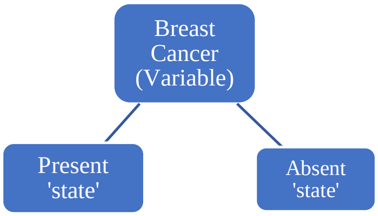

Figure 1: Possible states of node in Bayesian Network

F. Decision Tree

Decision Tree [16] is a powerful classification algorithm that is used to build a binary tree amongst the features available in datasets. The popular algorithms are ID3, C4.5, C5, J48, CART etc. The method of choosing the root node is very crucial. For this Decision Tree uses some mathematical approaches such as Entropy, Information Gain, Gini Index, Chi-Squared Test etc. to identify the variable in order to construct a decision-making tree. In doing so, Homogeneity order is to be maintained while distributing the variable into subsets.

G. Naïve Bayes

This classifier [10,11] is based on conditional probability. The attribute available in dataset is considered as strong and independent of each other. It uses less no. of parameters to make more effective. This classifier can be used in spam detection, language detection, sentiment analysis.

IV. METHODOLOGIES USED

Following figure 2 depicts the work flow in the direction to find the level of accuracy of the cancerous dataset using ML techniques.

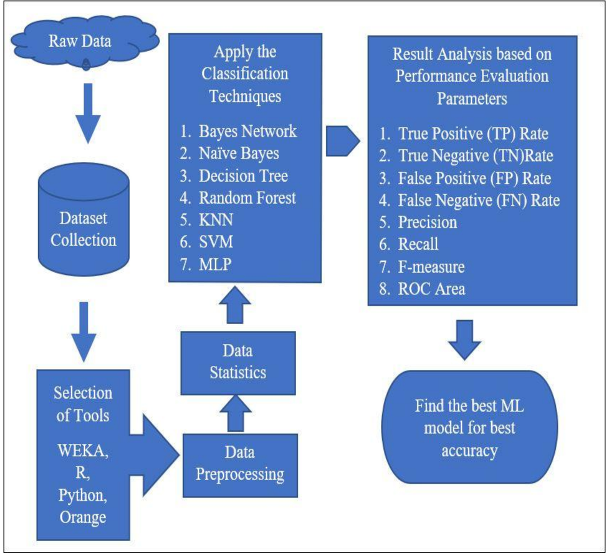
Raw Data
Dataset Collection
Selection of Tools
WEKA,
R,
Python,
Orange
Data Preprocessing
Data Statistics
Apply the Classification Techniques
1. Bayes Network
2. Naïve Bayes
3. Decision Tree
4. Random Forest
5. KNN
6. SVM
7. MLP
Result Analysis based on Performance Evaluation Parameters
1. True Positive (TP) Rate
2. True Negative (TN)Rate
3. False Positive (FP) Rate
4. False Negative (FN) Rate
5. Precision
6. Recall
7. F-measure
8. ROC Area
Find the best ML model for best accuracy

Figure 2: Sequence of steps for data classification

Some key terms used in figure are explained below**.**

Raw Dataset - There are huge number of raw dataset available on the Internet. Datasets are trusted and informative also. These datasets are available freely and publicly for the academic purpose. One of the resource of data is UCI Machine Learning.

Dataset Collection - The data are collected from reputed website UCI Machine Learning. There are two sets of datasets. The one dataset which has 11 attributes along with 699 instances and the second one is collected from same resource having 32 attributes along with 569 instances.

Selection of Tools – There are many open source tools available for the mining the data. E.g. WEKA, R-programming, Orange, KNIME, NLTK and the chosen tool is here WEKA.

Data Pre-processing - WEKA tool is able to pre-process the data in numeric and nominal (char) type. Both the dataset belong to breast cancer and the format of dataset is comma separated value (csv). The comma separated value (.csv) is useful for data analysis but required to change it to attribute-relation file format (.arff) for WEKA tool.

Data Statistics - Data is divided into labelled class to signify the distinct classes in the bar graph shown in figure 3 & 4 for the respective datasets. WEKA also gives the result for number of instances, missing value and unique instances of the dataset.

Classification Techniques- There are large number of classification techniques available in WEKA tool. The main and popular algorithm has been taken to analyse the data and predict the accuracy of being cancerous or not. The algorithms are Bayesian Network, Naïve Bayes, Support Vector Machine, Multilayer Perceptron, k-Nearest Neighbour, Decision Tree and Random Forest.

Performance Evaluation - The accuracy of the algorithm is determined on the basis of various parameters such as TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area etc. each parameter is discussed below.

True Positive (TP) Rate - True Positive is the number of people who actually suffer from 'cancer' among those who were diagnosed 'cancerous' [4]. It is also known as 'Correctly Identified'. TP Rate is the true positive instances correctly identified in a given class.

True Negative (TN) Rate - True Negative is the number of people who are 'non-cancerous' among those who were diagnosed 'cancerous' [4]. It is also known as 'Incorrectly Identified'. TN rate is the true negative instances incorrectly identified in a given class.

False Positive (FP) Rate - False Positive is the number of people who are 'cancerous' but were diagnosed as 'non-cancerous' [4]. It is also known as 'Correctly Rejected'. FP rate is the rate of false positive instances falsely classified in a given class.

False Negative (FN) Rate - (Incorrectly Rejected) False Negative is the number of people found to be 'non-cancerous' among those who ere diagnosed as 'cancerous' [4]. It is also known as 'Incorrectly Rejected'. FN rate is the rate of false negative instances falsely classified in a given class.

Precision – It is the proportion of instances that are truly of a class divided by the total instances classified as that class

Recall – It is the proportion of instances classified as a given class divided by the actual total in that class. It is equivalent to TP rate.

F-measure – It is also known as F-score, A combined measure for the precision and recall, calculated as
F-Measure = (2* precision*Recall)/(Precision + Recall)
ROC - It stands for Receiving Operational Curve, following three parameters are used for it.

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = (TP)/(TP+FN)
Specificity = (TN)/(TN+FP)

V. DATA ANALYSIS & RESULT

Dataset is obtained from UCI Machine Learning repository [www.ics.uci.edu]. There are two sets of breast cancer datasets shown in table 1 along with number of features, number of instances and missing values. Both the datasets have different features. The experiments are conducted using WEKA 3.8.2 tool on Window 10 platform.

Table 1: Description of Input parameters

Dataset	**No. of columns**	**No. of Instances**	**Missing values**
Breast Cancer Wisconsin Dataset (*BCWD11)	11 (Actual 10 Features)	699	16
Wisconsin Breast Cancer Dataset (**WBCD32)	32 (Actual 30 Features)	569	None

*Breast Cancer Wisconsin (BCWD11) dataset was obtained from the University of Wisconsin Hospitals, Madison USA. The above dataset was ready for the experiments from 15 July 1992 with 11 features.

1. Id number
2. Clump Thickness
3. Uniformity of Cell Size
4. Uniformity of Cell Shape
5. Marginal Adhesion
6. Single Epithelial Cell Size
7. Bare Nuclei
8. Bland Chromatin
9. Normal Nucleoli
10. Mitoses
11. Class

**Wisconsin Breast Cancer Dataset (WBCD32) was obtained from the University of Wisconsin Hospitals, Madison USA. The dataset was made available from November 1995 with 32 features.

1. Id
2. Diagnosis
3. Radius_mean
4. Texture_mean
5. Perimeter_mean
6. Area_mean
7. Smoothness_mean
8. Compactness_mean
9. Concavity_mean
10. Concave points_mean
11. Symmetry_mean
12. Fractal_dimension_mean
13. Radius_se
14. Texture_se
15. Perimeter_se
16. Area_se
17. Smoothness_se
18. Compactness_se
19. Concavity_se
20. Concave points_se
21. Symmetry_se
22. Fractal_dimension_se
23. Radius_worst
24. Texture_worst
25. Perimeter_worst
26. Area_worst
27. Smoothness_worst
28. Compactness_worst
29. Concavity_worst
30. Concave points_worst
31. Symmetry_worst
32. Fractional_dimension_worst

To maintain the consistency in the datasets, the missing value has been replaced by the mean value of the respective attributes.

In the other dataset of breast cancer named WBCD, no missing value is found and hence the experiment has been performed, obtained results are given below.

A. Pre-Processing on WEKA

First of all, the obtained dataset has been pre-processed on WEKA tool. It shows various types of results such as relation of attributes, number of instances, number of attributes. The dataset we retrieved was in the format of .csv file format (Comma separated value). First the file format had to change to make it compatible with WEKA. In the beginning of the pre-processing, the dataset had to be in the format of .arff file format (Attribute Relation File Format) to be explored by WEKA explorer. Finally uploading the compatible dataset file is done under pre-process tab in WEKA Explorer.

When the file is loaded. The descriptions of file come up and describe the important components such as current relation of the attributes, list of attributes, selected attributes along with Missing value, Distinct value, Unique value and Type of attributes. In addition to this, WEKA also calculate the Minimum, Maximum, Mean and Standard Deviation (std. dev.) of the data.

Figure 3 shows the graph of distribution of class as 'Benign' and 'Malignant' for the dataset *BCWD11 having no. of instances 699 along with 11 features.

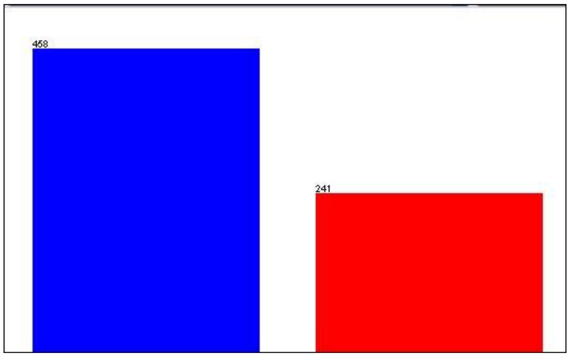

Figure 3 :*BCWD11 Class distribution (Benign-458 & Malignant-241)

Figure 4 shows the graph of distribution of class as 'Benign' and 'Malignant' for the dataset **WBCD32 dataset having no. of instances 569 along with 32 features.

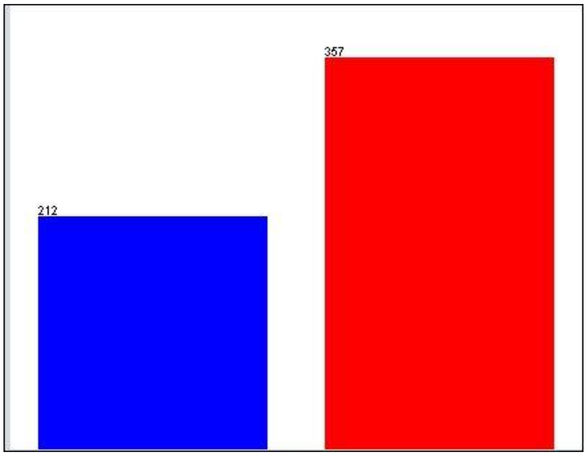

Figure 4: **WBCD32 dataset class Malignant-212 & Benign-357

B. Classification of attributes on WEKA

Both the datasets are classified by seven Machine Learning classifier algorithm. All are performed at 10-fold cross validations. These algorithms are Bayes Net, Naïve Bayes, SVM, Multilayer perceptron, k-NN, Random Forest and Decision Tree (J48).

Comparison of these algorithms is shown below in table 2 in terms of instance classified correctly and incorrectly followed by time taken to build model.

Table 2: Comparison of ML Algorithms on *BCWD11 dataset

Algorithm	**Correctly classified instances**	**Incorrectly classified instances**	**Time taken to build model**
Bayes Net	679	20	0.01 sec
Naïve Bayes	**671**	**28**	**0 sec**
(SVM)	676	23	0.01 sec
MLP	670	29	0.66 sec
KNN (IBK)	**665**	**34**	**0 sec**
Decision Tree (J48)	661	38	0.03 sec
Random Forest	674	25	0.13 sec

In order to build the model for the dataset BCWD11, Naïve Bayes and KNN classifiers take equal time to classify instances. Consideration for the building model for both the datasets is shown in table 3.

*Table 3: Comparison of ML Algorithms on **WBCD32 dataset*

Algorithm	**Correctly classified instances**	**Incorrectly classified instances**	**Time taken to build model**
Bayes Net	542	27	0.01 sec
Naïve Bayes	**527**	**42**	**0 sec**
SVM	557	12	0.01 sec
MLP	547	22	2.32 sec
KNN (IBK)	**547**	**22**	**0 sec**
Decision Tree (J48)	529	40	0.01 sec
Random Forest	546	23	0.15 sec

In order to build model for the dataset WBCD32, Naïve Bayes and KNN classifiers take equal time to classify instances. Consideration for the building model is both.

C. Performance metrics

This section evaluates the experimental results with precision, recall, F-measure, and ROC area for weighted average.

Table 4: Comparison of Performance Parameters for *BCWD11

Algorithm	**Performance Parameters**			
	Precision	**Recall**	**F-Measure**	**ROC Area**
BayesNet	**97.2%**	**97.1%**	**97.2%**	**99.2%**
Naïve Bayes	96.2%	96%	96%	98.6%
SVM (SMO)	96.7%	96.7%	96.7%	96.5%
MLP	95.9%	95.9%	95.9%	98.9%
KNN	95.1%	95.1%	95.1%	94.5%

(IBK)				
Decision Tree (J48)	94.6%	94.6%	94.6%	95.5%
Random Forest	96.4%	96.4%	96.4%	99%

The performance parameter measurement in table 4 gives a very promising result by BayesNet classifier in case of BCWD11 dataset. This classifier reaches 99.2% for ROC area, 97.2% for F-Score, 97.1% for Recall and 97.2% for Precision.

Table 5: Comparison of Performance Parameters for **WBCD32

Algorithm	**Performance Parameters**			
	Precision	**Recall**	**F-Measure**	**ROC Area**
BayesNet	95.3%	95.3%	95.3%	98.4%
Naïve Bayes	92.6%	92.6%	92.6%	97.6%
SVM (SMO)	**97.9%**	**97.9%**	**97.9%**	97.3%
MLP	96.1%	96.1%	96.1%	99.1%
KNN (IBK)	96.1%	96.1%	96.1%	95.6%
Decision Tree (J48)	93%	93%	93%	92.3%
Random Forest	96%	96%	95.9%	**99.1%**

The best model is chosen as SVM for WBCD32 dataset shown in table 5 where all performance parameters approaching 97.9% except ROC area 99.1% by Random Forest.

D. Level of accuracy of ML Techniques on WEKA

Now it is time to find the level accuracy of the ML algorithm performed on two datasets. Which ML algorithm performed better towards the dataset.

Table 6: Comparison of classification algorithm

Algorithm	**Accuracy Percentage**	
	***BCDW11 dataset**	****WBCD32 dataset**
BayesNet	**97.13%**	95.25%
Naïve Bayes	95.99%	92.61%
SVM (SMO)	96.7%	**97.89%**
MLP	95.85%	96.13%
KNN (IBK)	95.13%	96.13%
Decision Tree (J48)	94.56%	92.97%
Random Forest	96.42%	95.95%

The comparison of ML algorithms against two datasets of breast cancer is shown in the table 6. The BayesNet performed more accurate when the features of dataset are comparatively less and Support Vector Machine (SVM) model performed well when the feature of dataset is comparatively more. The number of samples (instances) also has a major significance when accuracy is to be measured.

The Line chart of the above table 6 is shown in figure 5 for the accuracy predicted on the dataset *BCWD11 and **WBCD32 respectively.

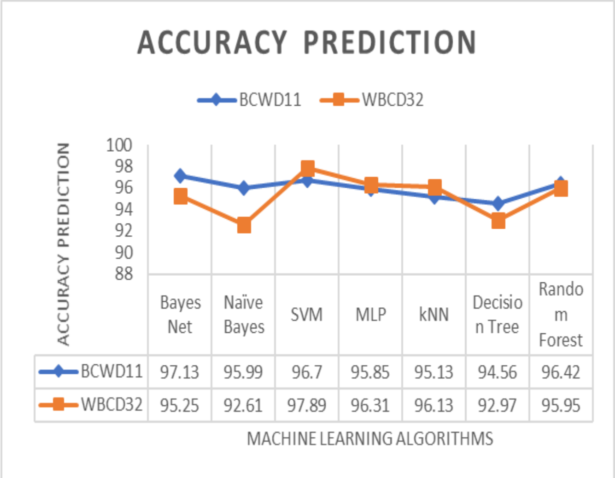

Figure 5: Prediction of accuracy (in percentage) in Line Chart

Results shown in the line chart, the best outcome is produced by Bayesian Network and SVM on the dataset *BCDW11 and **WBCD32 respectively. The other classifiers are also performed well in estimating and predicting the accuracy, additionally experiment performed on the given datasets shows that number of features matters a lot in the dataset.

VI. CONCLUSION & FUTURE WORK

The experiments were performed on two datasets of breast cancer, taken from UCI Machine Learning repository. These two datasets have different features, with 11 features and with 32 features. Seven Machine Learning algorithms are applied on both the datasets. The prediction of cancerous elements is likely to be found in the both the datasets. The datasets are divided into two parts. One part is called training data which is 65% of total dataset, and rest remaining 35% is the test data. In case of BCDW11 dataset, Bayesian Network ML technique produced the accuracy of 97.13% while for WBCD32 dataset, Support Vector Machine (SVM) ML technique gave the accuracy of 97.89%.

Bayesian Network ML technique gives best accuracy with less featured data while Support Vector Machine ML technique gives best accuracy for more features data.

In future, we intend to use the same dataset with dependant features only, then effect of it will be observed on accuracy level of the results for cancer prediction.

REFERENCES

[1] B. Nithya, V. Ilango, 2017, "*Comparative Analysis of Classification Methods in R Environment with two Different Datasets*.", Intl J Scientific Research and Computer Science, Engineering and Information Technology (IJSRCSEIT), vol 2, Issue 6, ISSN: 2456-3307.

[2] D. Lavanya et al. 2011, *"Analysis of Feature Selection with Classification: Breast Cancer Datasets."*, Intl. J of Computer Science & Engineering (IJCSE), vol. 2, No. 5, Oct-Nov 2011, ISSN: 0976-5166.

[3] Deepika Verma et al., 2017. "*Analysis and Prediction of Breast Cancer and Diabetes disease datasets using Data Mining Classification Techniques."*, IEEE Xplore Proceeding of the Intl. Conf. on Intelligent Sustainable Systems (ICISS 2017), IEEE Xplore Compliant – part number: CFP17M19-ART, ISBN:978-1-5386-1959-9

[4] Niyati Gupta et al. 2013, "*Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare data.*", IOSR J. of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, vol.11, issue 5, pp 70-73, May-Jun 2017.

[5] Morteza Heidari et al. 2017. "*Prediction of Breast Cancer Risk using Machine Learning Approach embedded with a Locality Preserving Projection Algorithm.*" Institute of Physics in Medicine and Biology (IPEM), doi: https://doi.org/10.1088/1361-6560/aaa1ca.

[6] T. John Peter et al. 2012. "*Study and Development of Novel Feature Selection Framework for Heart Disease Prediction*." Intl J. Scientific and Research Publication, IJSRP. Vol.2 Issue 10, (Oct. 2012), ISSN: 2250-3153.

[7] Mahua Nandy , 2013. "*An Analytical study of Supervised and Unsupervised Classification Methods for Breast Cancer Diagnosis*". 2nd Intl conf on Computing Communication and Sensor Network (CCSN-2013), Proceedings published by Intl. J Computer Application (IJCA) .

[8] Wenbin Yue, Zidong Wang, Hongwei Chen, and Annette Payne. May 2018. "*Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis."*, www.mdpi.com/journal/designs Design 2018, 2, 13; doi:10.3390/designs2013.

[9] Mohd. Milon Islam et al. 2017. *Prediction of Breast Cancer using Support Vector Machine and K-Nearest Neighbours."*, 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dec 21-23, 2017, Dhaka, Bangladesh.

[10] Amit Bhola, Arvind Kumar Tiwari, December 2015, "*Machine Learning Based Approaches for Cancer Classification using Gene Expression Data.*", Machine Learning and Application: An Intl. J. (MLAIJ), Vol 2, No.3/4

[11] Pedro D.,Micheal P., 1997, "*On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.*", Machine Learning, 29, 103-130 (1997), Kluwer Academic Publishers, Netherlands.

[12] Wenbin Yue, Zidong Wang, Hongwei Chen, and Annette Payne. May 2018. "*Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis.*", www.mdpi.com/journal/designs Design 2018, 2, 13; doi:10.3390/designs2013,.

[13] Isabelle Guyon, Jason W., Stephen B., et al., 2002, "*Gene Selection for Cancer Classification using Support Vector Machine, Machine Learning,*" vol.46, pp 389-422, 2002, Springer, Kluwer Academic Publishers, Netherlands

[14] D. Coomans, D.L.Massart, 1982, "*Alternate k-Nearest Neighbour Rules in Supervised Pattern Recognition, Part-1. k-NN classification by using Alternative Voting Rules.*", Analytica Chimica Acta, 136 (1982) 15-27, Elsevier Scientific Publishing Company, Amsterdam, Netherlands

[15] Leo Breiman, 2001, "*Random Forests, Machine Learning"*, vol. 45, issue 1, pp 5-32, Oct 2001, Springer, Kluwer Academic Publishers, Netherlands

[16] Dursun Delen, Glenn Walker, Amit Kadam, 2005, "*Predicting Breast Cancer Survivability: a comparison of three data mining methods.*", ELSEVIER Artificial Intelligence in Medicine (2005), 34, 113-127. doi: 10.1016/j.artmed.2004.07.002

[17] A. Marcano-Cedeno et al. 2011, "*WBCD breast cancer database classification applying artificial metaplasticity neural network.*", ELSEVIER Expert Systems with Applications 38 (2011) 9573-9579. Doi: 10.1016/j.eswa.2011.01.167

[18] B.B. Chaudhauri, U. Bhattachrya, 2011, "*Efficient training and Improved Performance of Multilayer Perceptron in Pattern Classification*". ELSEVIER Neurocomputing 34 (2000) 11-27

[19] J. Tang, C. Deng, G. Huang, 2016, "*Extreme Learning Machine for Multilayer Perceptron"*, IEEE Transaction on Neural Networks and Learning System, vol 27, No. 4, April 2016.

[20] Cruz-Ramirez Nicandro et al., 2013, "*Evaluation of the Diagnostic Power of Thermography in Breast Cancer using Bayesian Network Classifiers.*", Hindwai Publishing Corp, Computational and Mathematical Methods in Medicine. Vol 2013, Article ID 264246, 10 pages, http://dx.doi.org/10.1155/2013/264246

[21] S. Wongthanavasu, 2010, "*A Bayesian Belief Network Model for Breast Cancer Diagnosis."*, Springer Operation Research Proceedings 2010, Intl. Conf. German Operation Research Society, Sept 1-3, 2010, pp 3-8,

Authors Profile

Mr. Ajay Kumar pursed Bachelor of Engineering from SJCE Mysore, affiliated to VTU Belagaum, Karnataka India in 2005 and Master of Engineering from BIT Mesra Ranchi, India in year 2007. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Sciences & Engineering, DIT University Dehradun, India. He is a member of CSI (computer society of India) since 2011, a life member of the ISTE, ACM since 2016. He has published more than 6 research papers in reputed international journals and conferences. His main research work focuses on Big Data Analytics, Data Mining, Machine Learning and Computational Intelligence based education. He has 2 years of industrial experience, 10 years of teaching experience and 4 years of Research Experience.

Prof. (Dr.) Rama Sushil currently working as Professor in Department of Information Technology, DIT University Dehradun since 2013. She has served as a head of the department for last 5 years at DIT. She has total teaching and research experinec more then 17 years. She has comlpleted PhD from IIT Roorkee, Uttarakhand , India. She has guided various PhD students. She has also authored chapters in the book of IGI Global and published more than 40 research papers. Her research interests are cloud computing, distributed computing, big data, machine learning.

Dr Arvind Kumar Tiwari received his B.E. degree in Computer Science & Engineering from CCS University, Meerut, India and M.Tech in CSE from UPTU, Lucknow and Ph.D. in CSE from IIT(BHU), Varanasi, India. He has worked as Professor and Vice Principal in GGS College of Modern Technology, Kharar, Punjab, India.and currently working as an Associate Professor in KNIT Sultanpur, U.P. India. He is a member of IEEE & IEEE computer society ACM. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE. His main research work focuses on Big Data Analytics, Computational Intelligence, Pattern Recognition. He has more then 10 years of teaching experience and 4 years of Research Experience.