Hyper-parameter tuning of the best classifier evaluated through comprehensive comparison of supervised learning algorithms

[1]Ms. Meenakshi V., [2] Ms. Aruna Devi P., [3] Dr.(Ms.) Chamundeeswari M.
[1]Assistant Professor,[2]Assistant Professor & Head,[3]Associate Professor
[1]Department of Computer Science,
[1]Lady Doak College, Madurai, India.

Abstract: Breast cancer is one of the major diseases which when diagnosed earlier and treated have a greater chance of recovery. This paper aims to predict the malignant cells against the benign cells taking a dataset with about 32 attributes of such cells. Machine learning helps us to build a computer system that can automatically learn and improve with experience without being explicitly programmed. It does not rely on rule-based programming rather it works on data and learns from experience. Classification is the most classical supervised machine learning algorithm. It helps in identifying the set of class labels for the new observations. The fundamental goal of classification problem is to interpret the data that is never seen before. In this paper, prediction of cancer is performed using four different machine learning algorithms such as Decision Tree, Neural Network, Support Vector Machines and Random Forest. Experimental results have been analysed using various performance measures and it has been concluded that the Random Forest Model performs better. Further tuning of hyper-parameters results in increased performance.

Index Terms **- Classification, Machine Learning, Hyper-parameter Tuning, Grid Search, Random Forest**

I. INTRODUCTION

Breast cancer has become the most common disease among women. Epidemiology of breast cancer across different Population Based Cancer Registries (PBCR) in India shows increasing trends for incidence and mortality mainly due to rapid urbanization, industrialization, population growth and ageing affecting almost all parts of India. Breast cancer has ranked number one cancer among Indian females with age adjusted rate as high as 25.8 per 100,000 women and mortality 12.7 per 100,000 women [17]. Earlier detection of breast cancer will help women in their survival. Diagnosis of the breast cancer involves the classification of benign and malignant cells for the medical practitioner.

Machine Learning [19] is now widely used to speed up the decision-making process by building the algorithm or model with the help of direct experience from the historical data. This data helps the model to learn and boost the prediction of unknown hidden facts. The accuracy of the prediction depends upon the newly developed model. Supervised machine learning algorithm helps to find a rule or a set of rules that classifies the data depending on the class label [16]. The vital role of machine learning is to accelerate the diagnosis of malignant cells. This helps in increased chance of the recovery by treating at an earlier stage.

Machine Learning classification algorithms like Decision Tree, Random Forest, SVM and Neural Network were implemented to create a model that could predict the target class with factors 'M' for Malignant and 'B' for Benign cells and then the performance of the best algorithm is found based on performance measures.

Further, the parameters for Random Forest algorithm have been tuned systematically to find the optimal value resulting in an increased accuracy.

II. BACKGROUND

Breast cancer is considered to be a global health problem. This was confirmed in the article that was published in Times of India during April 2018 that was written by Nithin Gangane of Mahatma Gandhi Institute of Medical Sciences where he quoted that breast cancer has suddenly become the number one cancer. He proved his articulate by conducting two cross sectional studies. The study concentrated on patient delay, system delay, quality of life and self-efficacy [10].

Lavanya D. and Rani D. K. U. [12] proposed a hybrid approach of combining CART with feature selection and bagging on breast cancer data set and found out that the hybrid approach works fine based on classification accuracy.

Mojarad S. A., Dlay S. S., Woo W. L. and Sherbet G. [15] explored the predictive potential of the markers in the state of breast cancer and the accuracy of the Neural Network is accessed by the use of stratified and k-fold validation. Scalable Conjugate Gradient algorithm is used for training the multi-layer perceptron and his work concluded that Neural Network is the best modelling approach for cancer diagnosis.

Chandra Prasetyo, AanKardiana and Rika Yuliwulandar [6] implemented the ANN with extreme learning which has a better generalization classifier model than back propagation Neural Network. Analysis has been done through 5-fold cross validation techniques with three runs and the accuracy has been found good for the extreme learning model.

Devendra Kumar Tiwary [7] performed a comparison of four machine learning algorithm such as Decision Tree, Naive Bayes, Artificial Neural Network and Support Vector Machine using WEKA on credit card fraud detection data set and observation conveys that Decision Tree algorithm is considered to be the best suited algorithm for this data set.

Md. Nurul Amin [14] presented in his paper a comparison of different classification on Hematological Data using WEKA. The methods used are J48, Multilayer perception and Naive Bayes and concluded that the J48 Decision Tree algorithm outperforms well on that dataset.

Thorsten Joachims [16] in his paper, performed Text Categorization with Support Vector Machines and concluded that SVM outperforms all the other algorithms. They are fully automatic and do not require tuning of parameters.

Leo Breiman [3] has stated that for recent problems like medical diagnosis and document retrieval have many input attributes with each one containing only a small amount of information. A single tree classifier will have accuracy only slightly better than a random choice of class. But combining trees grown using random features can produce improved accuracy.

III. STATE OF ART

Machine learning is the subset of Artificial Intelligence that uses data and applies statistical techniques to build an analytic model without being explicitly programmed. This analytic model is then used for predictive analysis and decision making. Classification is process of assigning class labels to instances, given a training set of classified examples. Here all the classification algorithms are implemented using R tool. All the features of the data set have been considered for classification, 70% of data set is used for training and 30% of data set is used for testing.

3.1 Decision Tree

Decision Tree is one of the most powerful classification algorithms for decision making and knowledge discovery that classifies the labeled training data into rules in tree form. It is a supervised learning algorithm used for both regression and classification. In the tree representation each internal node represents an attribute and each leaf node represents a class label thereby learning decision rules from the training data.

Decision Tree classification technique employs two phases: tree building and tree pruning. Tree building is achieved using top-down approach. It is during this phase that the tree is recursively partitioned till all the data items belong to one class label. The tree construction process uses entropy, a measure from information theory that characterizes the impurity. The higher the entropy lower the information gain.

Given a set of classes, C = {1, ……., m} with equal probability of occurrences, the entropy E is

$$E = -p_1 \log p_1 - p_2 \log p_2 - ..... - p_m \log p_m$$

where p_i is the probability of occurrence of i. The attribute with lowest entropy is selected as split criteria for the tree.

Tree pruning is done in a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting [9], [13]. Overfitting in Decision Tree algorithm may lead to misclassification error. Tree pruning is less complex compared to the tree growth phase as the training data set is scanned only once [2].

Here in our study we have implemented Decision Tree algorithm using rpart package. Some important parameters for this algorithm are 'method' and 'split' where method is class and the split function is the information gain.

3.2 Random Forest

Random Forest (RF) is an ensemble machine learning algorithm that is a combination of tree predictors where each tree depends on the values of a random vector which is sampled independently and for all trees in the forest with same distribution" [3]. This algorithm is advantageous over other techniques as it has the ability to handle highly non-linear biological data, robustness to noise and tuning simplicity yielding best accuracies.

The Random Forest algorithm is a collection of tree-structured classifiers:

$$f(x,\theta_k) , k=1,2...K$$

where θ_k is a random vector that meets i.i.d. (independent and identically distributed) assumption [5] and each tree casts a unit vote for the most popular class at input x. For classification problems, the forest prediction is the unweighted plurality of class votes (majority vote). The algorithm converges with a large enough number of trees.

Here we have implemented Random Forest algorithm using randomForest package with number of variables randomly sampled (mtry) as 2 and number of trees to grow (ntree) which is 300.

Initially the implementation is done using the default values for all the parameters and the evaluation measures tabulated. The hmeasure package computes and reports the H measure of classification performance, alongside most commonly used alternatives, including the AUC. The package also provides convenient plotting routines that yield insights into the differences and similarities between the various metrics [1]. Breast Cancer prediction demands more accuracy in the current day scenario. Hence the need for optimization.

Hyper-parameters that are also called Tuning parameters include mtry, ntree and maxnodes. ntree is the number of trees to grow. Larger the tree, it will be more computationally expensive to build models. mtry refers to how many variables we should select at a node split. The default value is p/3 for regression and sqrt(p) for classification where p is the number of columns. Smaller values of mtry may lead to overfitting. nodesize refers to number of observations needed in the terminal nodes. This parameter is directly related to tree depth. Higher the number, lower the tree depth. With lower tree depth, the tree might even fail to recognize useful signals from the data.

3.3 Support Vector Machine

Support Vector Machine(SVM), developed by Vapnik [4] was primarily intended for binary classification.

$$f(w,x)=w.x+b$$

The main objective is to determine the optimal hyperplane separating the two classes in a given dataset having input features $x \in Rp$ and labels $y \in \{-1, +1\}$. SVM learns by solving the constrained optimization problem. The Support Vector machine is a generalization of a simple and intuitive classifier called the maximal margin classifier. A hyperplane is a flat affine subspace of hyperplane dimension $p - 1$ [11]. For instance, in two dimensions, a hyperplane is a flat one-dimensional subspace called a line. In three dimensions, a hyperplane is a flat two-dimensional subspace which a plane. In $p > 3$ dimensions, it can be hard to visualize a hyperplane, but the notion of a $(p - 1)$-dimensional flat subspace still applies. In two dimensions, a hyperplane is defined by the equation,

$$\beta_0+\beta_1X_1+ \beta_2X_2=0$$

for parameters β_0, β_1, and β_2. When we say that above equation defines the hyper-plane then $X=\{X_1, X_2\}$ holds the point in the hyperplane. The above is the equation of line in two dimension hyperplane.

It is implemented using **ksvm** method with kernel function which is used in training as radial basis kernel. Fitting is done on output data by performing 3-fold cross validation.

3.4 Neural Network

Neural Network is a mathematical model based on biological Neural Networks. It is an interconnected group of artificial neurons that processes information using connectionist approach for computation. It is a robust system that changes its structure based on the information flow that may be external or internal through the hidden layers during the learning phase. It has various structures based on the type of input-output data and the most widely used structure is the multilayer perceptrons.

The following equation summarises the calculated output:

$$f(x, w) = \varphi(x .w) = \varphi(\sum_{i=1}^{p} x_i . w_i)$$

In the equation, variables x and w represent the input vector and weight vector of the neuron when there are p inputs into the neuron. Greek letter (phi) denotes an activation function. The process results in a single output from a neuron.

The implementation is done with **nnet** package in R. The model is trained with 10 units of hidden layer, linear output, trace optimization and skip layer connection is set to true. Maximum number of weights used by the model is 10000 and maximum iteration is 100.

IV. PERFORMANCE METRICS

Metrics is most important to evaluate our machine learning model. The performance of the machine learning algorithm is measured and compared using the metrics such as Precision, Recall/ Sensitivity (Sens), Specificity(Spec), AUC and Accuracy. All these measures make use of the values of the confusion matrix in table 4.1 as given below

Table 4.1: Confusion Matrix

		Actual	
Predicted		True Positive [TP]	False Positive [FP]
		False Negative [FN]	True Negative [TN]

4.1 Precision

Precision is the measure of number of correct classification for each class. The value of precision lies between 0 and 1. Precision value closer to 1 indicates maximum correct classification.

$$\text{Precision} = \frac{TP}{TP+FP}$$

4.2 AUC

Area under the curve (AUC) is calculated to measure the quality of classifier. The amount of area under the receiver operating characteristics (ROC) curve is AUC. The model scoring high AUC as compared to other models is considered as efficient model. Its value is between 0 and 1. The quality of model is good if it has AUC value near to 1.

4.3 Accuracy

Accuracy is calculated to measure the correctness of classifier. Accuracy can be calculated as:

$$\text{Accuracy} = \frac{TP+TN}{Total} * 100$$

4.4 Sensitivity/ Recall

Sensitivity(Sens) is also known as recall or true positive rate. It is the proportion of actual positives which are correctly identified as positives by the classifier and is computed as:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

4.5 Specificity

Specificity(Spec) is also known as true negative rate. It relates to the classifiers ability to identify negative results and is computed as:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

TN: True Negative, FP: False Positive, TP: True Positive and FN: False Negative

V. RESULTS & DISCUSSION

For the purpose of study, Wisconsin Diagnostic Breast Cancer (WDBC) dataset was taken from the UCI repository [8]. Features of the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Fig.1 depicts benign cells whereas Fig.2 depicts malignant cells [20,21].

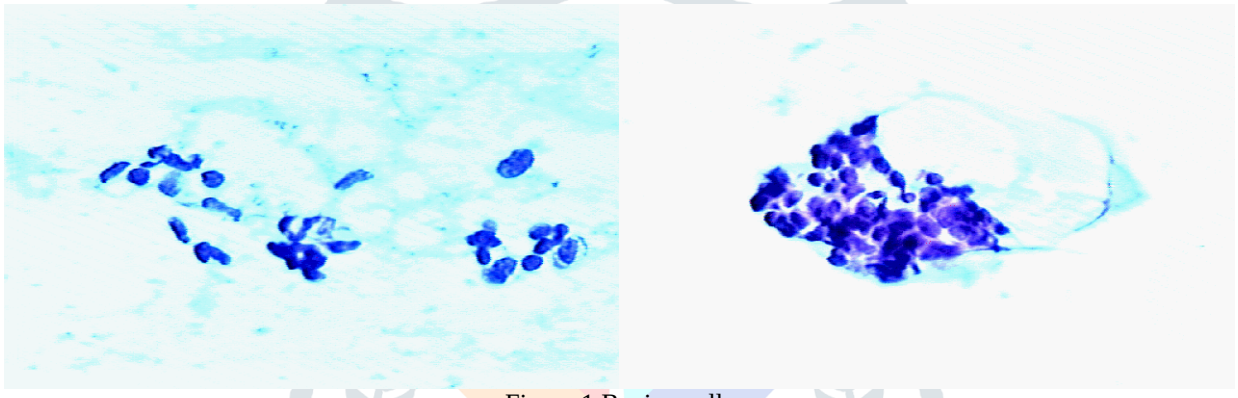

Figure 1 Benign cells

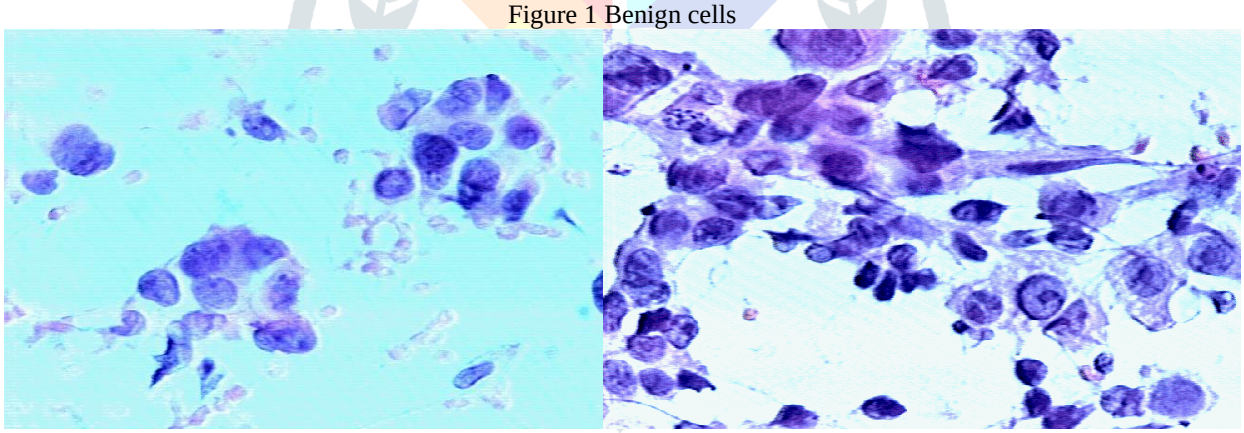

Figure 2 Malignant cells

The Dataset comprises 569 instances and 32 attributes that include ID, Diagnosis and ten real valued attributes computed for each cell nucleus. Other features are obtained by calculating the mean, standard error and worst or largest of these values computed for each image. Diagnosis is the target class that can take any one of the two values either M (Malignant) or B (Benign). As our machine learning algorithm can only read numerical values, it is essential to encode the categorical features to numerical values. Hence the target class is encoded to 1s and 0s shown in Table 5.1 which is the first step in preprocessing of data set.

Table 5.1: Target Class encoding of Breast Cancer Dataset

Categorical Feature	Encoded Numerical value
B	0
M	1

Here in our study, four machine learning algorithms are evaluated based on the above-mentioned metrics. To begin with, sensitivity is calculated for each one of them and analysed. It is noted that sensitivity or the true positive rate is

maximum for Random Forest than the other algorithms that are considered equally good from our prior study. The results are shown in Table 5.2 and Fig. 3.

Table 5.2: Sensitivity Measure for Performance Evaluation

Sensitivity	DT	RF	SVM	NN
Run 1	0.917	0.969	0.906	0.926
Run 2	0.957	0.949	0.956	0.929
Run 3	0.883	0.94	0.925	0.789
Run 4	0.879	0.88	0.873	0.81
Run 5	0.852	0.922	0.923	0.894

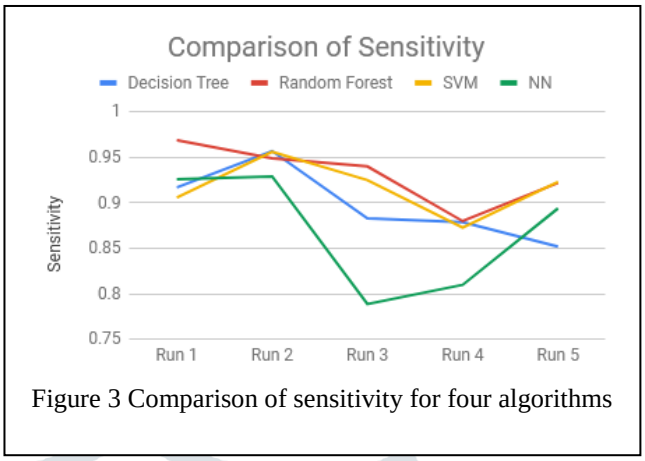

Figure 3 Comparison of sensitivity for four algorithms

Next, we consider the Specificity or True negative rate for each algorithm and results are depicted in Table 5.3 and Fig. 4. This clearly infers that Support Vector Machines and Neural Network algorithm perform equally good as Random Forest.

Table 5.3: Specificity Measure for Performance Evaluation

Specificity	DT	RF	SVM	NN
Run 1	0.939	0.991	0.992	1
Run 2	0.951	0.964	1	0.983
Run 3	0.919	0.99	0.971	1
Run 4	0.933	0.992	1	0.989
Run 5	0.918	1	1	0.976

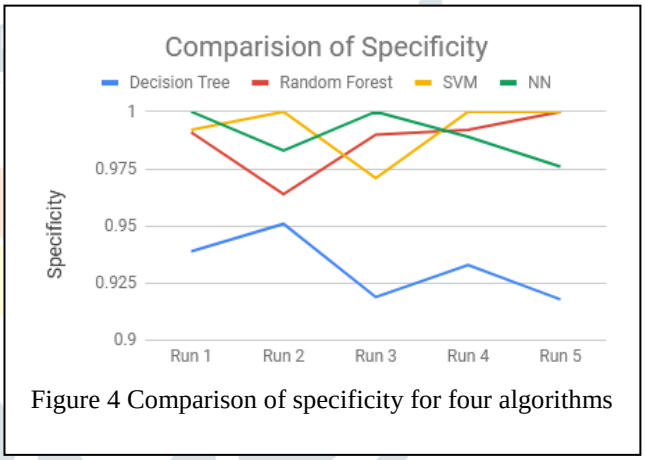

Figure 4 Comparison of specificity for four algorithms

Accuracy is a widely used metric that is finally tabulated to decide on the best performing algorithm. It is found that on an average of five runs, the Random Forest algorithm outperforms with an accuracy of 96.9 as depicted in Table 5.4 and Fig. 5

Table 5.4: Accuracy as a Performance

Accuracy	DT	RF	SVM	NN
Run 1	92.98	98.25	96.49	96.49
Run 2	95.32	95.91	98.25	94.74
Run 3	90.64	97.08	95.32	90.06
Run 4	91.23	95.91	95.32	88.89
Run 5	89.47	97.08	97.08	95.32

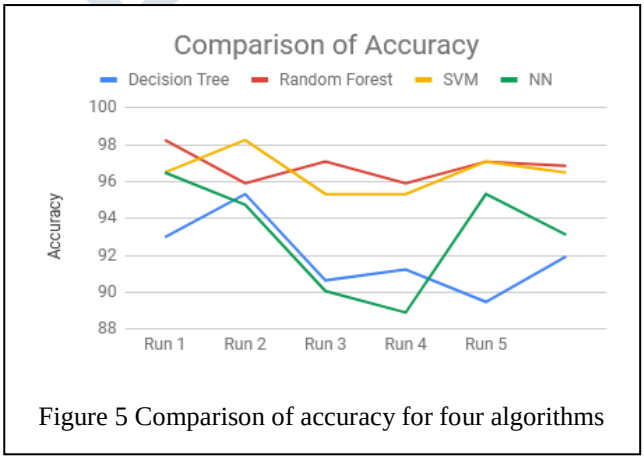

Figure 5 Comparison of accuracy for four algorithms

We also calculated the number of wrongly predicted instances as this is a very important factor when implementing in a medical diagnosis dataset. This is done with the confusion matrix obtained for each of the algorithms and is shown in Table 5.5 and Fig. 6.

Number of wrongly predicted instances= FP + FN

Table 5.5: Number of wrongly predicted instances

Wrongly predicted	DT	RF	SVM	NN
Run 1	12	3	6	5
Run 2	8	7	3	6
Run 3	16	5	8	16
Run 4	15	7	8	16
Run 5	18	5	5	8

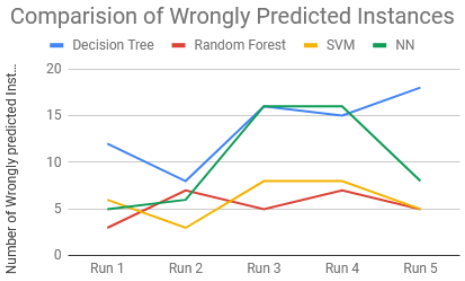

Figure 6 Comparison of Wrongly predicted Instances

It is found that Sensitivity and Accuracy is maximum for Random Forest algorithm, when considering Specificity SVM, Neural Network and Random Forest equally perform well. When we are concerned with the prediction of Malignant cells, it is very essential that the False positive and False negative predictions should be considerably low. Hence Random Forest again outperforms the rest of the algorithms. Therefore, it is concluded that Random Forest algorithm performs well among the four machine learning algorithms by building corresponding models and evaluating their performance. Further improvement of the accuracy is done by tuning the hyper-parameters of the Random Forest algorithm.

A lot of possible combinations of parameters are possible for the Random Forest algorithm. This can be done manually. Any way it can also be done by the machine using the Grid search method. In the Grid search method, the model will be evaluated for all the combinations that are passed in the function, using cross-validation.

To select the optimal model, the one with the minimum RMSE value is chosen. Hence by applying Grid search with cross validation folds, it is found that the mtry value 16 gives better accuracy. The mtry value 16 is now fine tuned to find the best mtry using the train() method of the caret package. Now this yields 15 as the best mtry value. ntree value has also been fine tuned to 210.In our efforts to further tune the max nodes hyper-parameter, the optimum value for maxnodes has been found to be 17. The evaluation parameters have been tabulated in Table 5.6.

Table 5.6: Performance measure on tuning

Evaluation Measures	Default	mtry 16	max nodes 17	Evaluation Measures	Default	mtry 16	max nodes 17
H	0.959	0.959	0.965	Precision	0.867	0.886	0.907
Gini	0.998	0.997	0.998	Recall	1	1	1
AUC	0.999	0.999	0.999	TPR	1	1	1
AUCH	0.999	0.999	0.999	FPR	0.045	0.038	0.03
KS	0.97	0.97	0.977	F	0.929	0.94	0.951
MER	0.012	0.012	0.012	Youden	0.955	0.962	0.97
MWL	0.011	0.011	0.008	TP	39	39	39
Spec.Sens95	0.985	0.977	0.977	FP	6	5	4
Sens.Spec95	1	1	1	TN	126	127	128
ER	0.035	0.029	0.023	FN	0	0	0
Sens	1	1	1	Accuracy	96.49	97.08	97.66
Spec	0.955	0.962	0.97	Total Time	6.94	52.15	119.29

VI. CONCLUSION

Our research study focused on classification of breast cancer using four supervised machine learning algorithms. Sensitivity, specificity, accuracy and number of wrongly classified instances are analysed and found out that Random Forest outperforms the others. In an attempt to further improve the performance of the model the parameters of the Random Forest algorithm have been fine tuned. Even though this tuning can be performed manually, Grid search with cross- validation has been implemented to ensure that optimal values are identified for each parameter.

REFERENCES

[1] Anagnostopoulos, C., canagnos@imperial.ac.uk, D.J. Hand, d.j.hand@imperial.ac.uk, N.M. Adams .adams@imperial.ac.uk, Measuring classification performance: the hmeasure package. Department of Mathematics, South Kensington Campus, Imperial College London, London SW7 2AZ September 10, 2012

[2] Bauer, E. and Kohavi, R. 2004. An empirical comparison of voting classification algorithms: Bagging, Boosting, And Variants, Machine Learning. 36(1-2), pp 105-139.

[3] Breiman, L. 2001. Random Forests. Machine Learning 45 (1), 5–32.

[4] Cortes, C. and Vapnik, V. 1995. Support-vector Networks. Machine Learning 20.3(1995), 273–297. Available from: https: //doi.org/10.1007/BF00994018

[5] Cover, T.M. and Thomas, J.A. 2006. Elements of Information Theory second edition. Wiley Interscience.

[6] Chandra Prasetyo, AanKardiana, Rika Yuliwulandari and Utomo. 2014. Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Machine Learning Techniques. International Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 7.

[7] Devendra Kumar Tiwary. March 2014. A Comparative Study of Classification Algorithms For Credit Card Approval Using Weka. International Interdisciplinary Research Journal, Vol.2 (3).

[8] Dua, D. and Karra Taniskidou E. 2017. UCI Machine Learning Repository. Available from: http://archive.ics.uci.edu/ml.https://archive.ics.uci.edu/ml/citation policy.html

[9] Forbes A. D. 1995. Classification Algorithm evaluation: Five performance measures based on confusion matrices, vol. 11, pp 189-206. Available from: https://link.springer.com/article/10.1007/BF01617722

[10] Gangane, Nitin. 2018. Breast Cancer in Rural India: Knowledge, Attitudes, Practices; Delays to Care and Quality of Life. Available from:http://umu.diva-portal.org/smash/get/diva2:1187627/FULLTEXT02.pdf

[11] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. An Introduction to Statistical Machine Learning: with Applications in R, Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 9, © Springer.

[12] Lavanya, D. and Rani, D. K. U. 2012. Ensemble Decision Tree Classifier for Breast Cancer Data. International Journal of Information Technology Convergence and Services (IJITCS) Vol, 2.

[13] Mansour, Y. 1997. Pessimistic decision tree pruning based on tree size. In Press of Proc. 14th International Conference on Machine Learning, pp.195-201.

[14] Md. Nurul Amin and Md. Ahsan Habib. Comparison of Different Classification Techniques Using WEKA for Hematological Data. American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-4, Issue-3, pp-55-61.

[15] Mojarad, S. A., Dlay, S. S., Woo, W. L. and Sherbet, G. 2010. Breast Cancer prediction and cross validation using multilayer perceptron Neural Networks. Paper presented at the Communication Systems Networks and Digital Signal Processing (CSNDSP) 7th International Symposium.

[16] Shawkat Ali, A. B. M. and Saleh A. Wasimi. 2009.Data Mining: Methods and Techniques, CENGAGE Machine Learning, India.

[17] ShreshthaMalvia, Sarangadhara Appalaraju Bagadi, Uma S. Dubeyand Sunita Saxena.Sept.2017. Epidemiology of Breast Cancer in Indian Women. Asia-Pacific Journal of Clinical Oncology, vol. 13, no. 4, pp. 289–295., doi:10.1111/ajco.12661. Available from: https://www.researchgate.net/publication/313545712

[18] Thorsten Joachims.1998. Text Categorization with Support Vector Machines: Machine Learning with Many Relevant Features.

[19] Tom Mitchell. 1997. Machine Learning.McGraw Hill. p. 2. ISBN 0-07-042807-7.

[20] W. H. Wolberg, W. N. Street and O. L. Mangasarian , Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine Needle Aspirates", journal = "Cancer Letters", year = 1994, volume = 77, pages = "163--171"

[21] www.cs.wisc.edu/~olvi/uwmp/cancer.html (images)