International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 03, March-2019

AUTOMATED ARTIFICIAL INTELLIGENCE BASED BREAST CANCER AND DIABETES PREDICTION

[1]Tawseef Ayoub Shaikh, [2]Rashid Ali

[1,2] *Department of Computer Engineering, Aligarh Muslim University, Uttar Pradesh, India*

Abstract: ***This paper carries a rigorous and in depth evaluation of top most four data mining algorithm techniques i.e. Naive Bayes, J48, SVM (Support Vector Machine) and an Ensemble in the form of Bagging, using ROC (receiver operating characteristics) as parameter. All four algorithms are evaluated on two benchmark datasets of breast cancer and diabetes and data processing is accordingly done in three different most extensively used data mining tools like Weka, Tanagra and MATLAB. The ROC in MATLAB always leaded the team in breast cancer case with an ROC value of 99.18 in all the four data mining algorithms case. This was followed by Bagging in Weka with an ROC value of 98.72, followed by Naive Bayes in Weka with ROC value of 97.54. Tanagra got the 96.18 as the highest ROC in Ensemble case. Both Weka and Tanagra got lesser values than the MATALB in all the four cases. On the other hand, in case of diabetes dataset, the trend changed where it is Naive Bayes with a value of 81.86 which topped the list in Weka, Tanagre and MATLAB. Similarly, in Tanagra case J48 with a value of 79.17 got the credit of leader. Again in SVM, MATLAB topped with a value 83.84 and finally in Ensemble, MATLAB with a value of 99.18 became the leader.***

Keywords: Navie Bayes, Ensemble, Data mining Algorithms, ROC (Receiver operating Characteristics), Weka, Tanagra

I. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer, which alone accounts for 30% all new cancer diagnoses for women, posing a threat to women's health [1]. According to the American Cancer Society (ACS), an estimation of 246,660 women will be diagnosed with breast cancer and approximately 40,450 women will die from this disease in 2016 [2]. In China, there has been an estimated of 214,360 women has died from breast cancer by 2008, and the number of death will reach up to 2.5 million by 2021 [3]. However, according to the survey, more than 30% of cancer cases will be surviving for a long-time if they accept the accurate early detection [4]. More than 3.5 million Americans are living with breast cancer, of whom 41,070 (40,610 women and 460 men) died from the disease in 2017. The chance of any woman dying from breast cancer is around 1 in 37, or 2.7 percent [5]. The diagnosis and treatment of breast cancer in the early stage is crucial to reduce the morbidity rates and prevent the progression of the disease. Globally, breast cancer comprises approximately 15% percent of all cancers affecting females [5]. Approximately 1 in 37 breast cancer patients will die as a result of the disease and it has been cited as the second most common cause of cancer related death amongst females. Breast cancer can occur in females of any age, but most commonly tends to affect females between the ages of 15 and 54 years old. In 2016, about 29% deaths were accounted in female due to breast cancer in the United State. In 2016, it was estimated that 595,690 American will die from cancer corresponding to 1,600 deaths per day. Moreover, the research evidences indicate that radiologists may miss up to 30 % of breast cancer depending up on the density of breasts.

Segmentation of breast images into functional tissues can aid tumor localization, breast density measurement, and assessment of treatment response, which is important to the clinical diagnosis of breast cancer. However, manually segmenting the images, which is skill and experience dependent, would lead to a subjective diagnosis; in addition, it is time-consuming for radiologists to review hundreds of clinical images. The average diagnostic accuracy of pathologists is approximately 75%. Also due to the similar clinical manifestations of different types of cancer, which make further analysis of the data very difficult.

Therefore, it has become necessary to automate some of the tasks in the diagnostic workflow to reduce the burden on the radiologist and pathologist. Fortunately, the development of computer vision and machine learning In this context, the machine learning techniques emerged as a dominant paradigm which offers reliable solutions and can perform some diagnostic task automatically and intelligently [6]. It has become the leading cause of cancer mortality among women younger than 45 years old [7] Breast density is often taken as a predictor of breast cancer risk assessment and prevention [8]. The percentage of breast density is calculated by dividing the area of the fibro glandular tissue by the total area of the breast. The odds ratio of developing breast cancer for women with most dense breasts is 2 to 6- fold greater than women with normal breast density [9]. Breast cancer is one of the leading causes of the high mortality rate in women [10]. Therefore, early detection is essential to lead an easy treatment and to increase the chances of survival as well [11]. The techniques used in

screening and monitoring of the breast cancer, including mammography, magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), thermography, and surgical incision [12–13]. Interpretation of the data obtained from these techniques is very complicated.

II. STATE OF THE ART STUDIES

Since automatic segmentation of breast images into functional tissues has received attention in recent years, amidst the more numerous studies of detection and segmentation of masses. This section focusses on recent important contribution in the same field as literature review.

In recent years, with the development of artificial intelligence, more and more data-driven intelligent classification approaches have been applied for breast cancer diagnosis, such as Naive Bayesian [14], Neural Network [15] Support Vector Machine (SVM) [16] or other hybrid algorithms [17, 18]. Na Liua at el.[19] put a novel breast cancer intelligent diagnosis approach, which employed information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection, The ranking of features is achieved according to IG algorithm, and extracting the top *m* optimal feature utilized the cost sensitive support vector machine (CSSVM) learning algorithm. To differentiate benign and malignant tissues, Veta et al. [20] mined the features of nuclei morphology in their work. Kowal et al. [21] aimed to segment nuclei using clustering algorithms for nuclei segmentation from biopsy microscopic images. This algorithm achieved high classification accuracy owing to a good feature extraction strategy in which morphological and texture features were employed. Similarly, Zhang et al. [22] used a set of parallel SVM classifiers and a set of artificial neural networks (ANN) to create a strong cascade classifier. The authors in at el. [23] demonstrated out of box techniques of transfer machine learning in comparison with the fully-trained network on the histo pathological imaging modality by considering three pre-trained networks: VGG16, VGG19, and ResNet50 and analyzed their behavior for magnification independent breast cancer classification. The results ranked fine-tuned pre-trained VGG16 with logistic regression classifier as best performer yielding 92.60% accuracy, 95.65% area under ROC curve (AUC), and 95.95% accuracy precision score (APS) for 90%–10% training–testing data splitting [23].

The much talked filed of machine learning i.e. deep leaning has also earned priority based attention nowadays Handsome work of deep learning has been done in breast cancer diagnosis. Wang et al. used sampling patches to train a CNN (Convolution neural network) to make patch-level predictions, then aggregated the results to create tumor probability heat maps and made slide-level predictions. The methodology was tested on the Camelyon16 dataset including 400 WSIs [24] with exceptional optimal results. The approach proposed in this paper is applied to the 4-class classification of breast cancer histology images and achieves 95% accuracy on the initial test set and 88.89% accuracy on the overall test set. The results are competitive compared to the results of other state-of-the-art methods. [25]. The authors in at el. [26] propose a convolutional neural networks (CNNs) for segmenting breast ultrasound images into four major tissues: skin, fibro glandular tissue, mass, and fatty tissue, on three-dimensional (3D) breast ultrasound images. Quantitative metrics for evaluation of segmentation results including Accuracy, Precision, Recall, and F1measure, all reached over 80%, which indicates that the method proposed has the capacity to distinguish functional tissues in breast ultrasound images.

Araújo et al. [27] proposed a CNN architecture designed for extracting features from patches of 512 * 512 pixels and performed 4-class classification based on 249 high resolution images released for the bio imaging 2015 breast cancer histology classification challenge. The results were very promising. Spanhol et al. [28] also proved using deep learning concepts that the feature extraction task was more difficult if the deep learning network was trained on breast tissue images at higher magnifications. Ciresan et al. [29] used convolutional neural networks (CNN) to detect mitosis in each of the patches extracted from stained breast biopsy slides. In [30], authors at el. proposed a novel genetic algorithm-based online gradient boosting (GAOGB) model for incremental breast cancer (BC) prognosis. The proposed GAOGB model is comprehensively evaluated on the U.S. National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program breast cancer dataset in terms of accuracy, area under the curve (AUC), sensitivity, specificity, retraining time, and variation at each iteration. Results show that the proposed GAOGB model achieves statistically outstanding online learning effectiveness.

The promising power of combining different but compatible algorithms in the form of ensemble has also been the core domain nowadays. Karianakis et al. [31] used the AdaBoost decision trees method to combine CNNs for classifying images. Gao et al. [32] developed a boosting algorithm that selected weak CNN learners among a set of CNNs to build a stronger learner for video recognition. In [33], authors at el. proposed the nested ensemble approach which used the Stacking and Vote (Voting) as the classifiers combination techniques as an ensemble methods for detecting the benign breast tumors from malignant cancers. The results demonstrate that the proposed two-layer nested ensemble models outperformance the single classifiers and most of the previous works. Both SV-BayesNet-3-MetaClassifier and SV-Naive Bayes-3-MetaClassifier achieved accuracy 98.07% (K = 10). However, SV-Naive Bayes-3-MetaClassifier is more efficiency as it needs less time to build the model.

III. DATASETS, MATRICES AND TOOLS

The benchmark datasets used for the evaluation of the selected data mining performance is the theme of this section. A brief section is diverted to evaluation matrices as well as the tools used in the work.

3.1 Data sets

In this study, two benchmark datasets in the form of breast cancer and diabetes are castoff in order to pinpoint the general best method and classifier.

3.1.1 Wisconsin breast cancer diagnosis (WBCD) The Wisconsin breast cancer diagnosis (WBCD) database [34, 35] developed by the University of Wisconsin Hospital grounded exclusively on an FNA test for breast masses finding. A total of 699 clinical instances is present in this dataset, possessing 458 (65.52%) benign and 241(34.48%) malignant. Every clinical instance is described by a set of 9 attributes having assigned integer values which are in range from 1 to 10 and one class has a binary value of either 2 or 4 as output as a convenience for representing benign and malignant cases respectively. Table 1(a) lists the physical meaning of the corresponding nine attributes. 16 occurrences are each absent one of the nine attributes among the 699 clinical. Frequent technique is to eliminate the cases where the data is not present. Removing artifacts and handling missing entries of datasets and addressing the issues with imbalanced datasets and proposing oversampling and under-sampling methods to handle this was done by the incorporation of PRTools (a Matlab integrated pattern recognition tool) using a set of linear, nonlinear and Bayes normal classifiers. From this dataset 16 missing occurrences are detached in order to gain high accuracy framing out the final dataset possessing 683 clinical occurrences, with 444 (65.01%) benign and 239 (34.99%) malignant cases.

Attribute	Values
Sample code number	Id number
Clump thickness	1-10
Uniformity of cell size	1-10
Marginal Adhesion	1-10
Single Epithelial Cell Size	1-10
Bare Nuclei	1-10
Mitoses	1-10
Class	2 and 4

S No.	Attribute	Type
1	**Number of times pregnant**	Numeric
2	**Plasma glucose concentration**	Numeric
3	**Blood pressure(Diastolic)**	Numeric
4	**Triceps skin fold thickness(mm)**	Numeric
5	**Numeric**	Numeric
6	**Body mass index(kg/m^2)**	Numeric
7	**Diabetes pedigree function**	Numeric
8	**Age (years)**	Numeric
9	**Class Variable (True or False)**	Nominal

Table 1: (a) Wisconsin breast cancer diagnosis (WBCD) (b) Diabetes dataset

3.1.2: Diabetes dataset Table 1(b) throws light on the diabetes dataset. It contains a total number of Instances as 768 with total number of attributes 8. All 8 attributes are numeric in nature except the last one which is a class label possessing a nominal data type [36]. Below is the representation of the dataset in terms of its conveyed meaning.

Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

3.2 Image descriptors

Features vector signifying the segmented region is mined as Quantitative measures (features) of different nature. Mean and standard deviation were utilized a descriptors for uncovering and classification of breast cancer formally, but higher order statistics like skewness and kurtosis has dominated the fashion nowadays. The 18 features mined embraces statistics (kurtosis, skewness, perimeter, area, minimum, maximum, standard deviation, mode and mean), shape (form, elongation, circularity, roughness) and texture (contrast, correlation, entropy, angular second moment, inverse difference moment), as depicted in Table 2 below. Wrapper method of dimensiolity reduction was prepared using the WEKA toolkit [37], dipping the dimensionality of each dataset to justifying 99% of its variability.

MATLAB [38, 39] was used for initial preprocessing and smoothing (with top-hat filtering and Gaussian low-pass) to mammograms so as to foil the loss of facts that could be valuable in later phases. Using Gaussian filter, a correlation kernel factor applied to the image was fashioned. Followed by the selection of the region of interest (ROI) and an image conversion into 4- bit, the first method of withdrawal of characteristics was applied. 150 findings interrelated to lesion tissue for feature quotation phase was acquired by this step. Gray-level co-occurrence matrix (GLCM) is a statistical scheme of texture investigation considering the spatial relationship between the image pixels is the first feature matrix used and the second matrix, the gray-level run length (GLRLM) which is grounded on computerizing the number of lines of grey levels at many angles. Weka [37, 41] and Tanagra [40] are both open source most cited data mining tools used most commonly in machine learning field. MATLAB [39, 41] also has inbuilt various toolboxes for carrying out image, signal, statistical, mathematical and machine learning works.

3.3 Classification Metrics

Consists a list of parameter normally used for finding out the classification accuracy of the classifier. All the evaluation parameters are calculated from the confusion matrix.

		Actual Outcome		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = Σ True Positive / Σ Test Outcome Positive
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = Σ True Negative / Σ Test Outcome Negative
		Sensitivity = Σ True Positive / Σ Condition Positive	**Specificity** = Σ True Negative / Σ Condition Negative	

Fig. 1: Confusion matrix

1): Sensitivity (also called *Recall sensitivity, recall, hit rate or true positive rate (TPR)*: Sensitivity is the proportion of actual positives which are correctly identified as positives by the classifier. [41]

Sensitivity =TPR= TP / (TP +FN) (1)

2): Specificity (also called *True Negative Rate*): Specificity relates to the classifier's ability to identify negative results. Consider the example of medical test used to identify a certain disease. The specificity of the test is the proportion of patients that do not to have the disease and will successfully test negative for it.

Feature	Description
Mean(i-mean)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi$, with *n* being the number of pixels inside the region delimited by the contour and *xi* being the grey level intensity of the i[th] pixel inside the contour.
Standard Deviation (i-std)	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
Skewness (i-skewness)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$
Kurtosis (i-kurosis)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$ -3
Energy(t-energ)	$\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$ while L being number of grey levels and p being the grey-level co-occurrence matrix and, thus, *p(i,j)* is the probability of pixels with grey-level *i* occur together to pixels with grey-level *j*.
Contrast (t-contr)	$\sum_{i}\sum_{j}(i-j)^2 p(i,j)$
Correlation (t-corr)	$\frac{\sum_{i}\sum_{j}(ij)p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$, with $\mu_x\mu_y$, $\sigma_x\sigma_y$ being the means and standard deviations of p_x and p_y, the partial probability density functions.

Table 3: Set of texture descriptors computed from the Grey-level

Specificity=TNR=TN / (TN+FP) (2)
3): Precision (positive predictive value (PPV)): This is a measure of retrieved instances that are relevant. In other words:
Precision=PPV= TP/ (TP+FP) (3)
4): Negative predictive value (NPV): NPV= TN / (TN+FN) (4)
5): Miss rate or false negative rate (FNR): FNR= FN/ FN+ TP=1-TPR (5)
6): Fall out or false positive rate (FPR): FPR= FP/ FP+ TN=1-TNR (6)
7): False Discovery rate (FDR): FDR= FP/ FP+ TP=1-PPV (7)
8): False omission rate (FOR) [41]: FOR= FN/ FN+ TN=1-NPV (8)
9): Accuracy: This is the simplest scoring measure. It calculates the proportion of correctly classified instances.
Accuracy = (TP + TN) / (TP+TN+FP+FN) (9)
10): F_1 score (also F-score or F-measure): The traditional F-measure or balanced F-score (F_1 score) is the harmonic mean of precision and recall:

$$F1= 2*\frac{1}{\frac{1}{recall}+\frac{1}{precision}} = 2*\frac{precision*recall}{precision+recall} \quad (10)$$

11): G-measure
While the F-measure is the harmonic mean of Recall and Precision, the G-measure is the geometric mean.

$$.G= \sqrt{Precision * \mathrm{Re}call} \quad (11)$$

This is also known as the Fowlkes–Mallows index.

12): Matthews correlation coefficient $$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (12)$$

13): Lift = (predicted rate / average rate) (13)
14): Brier score = (Actual result – Forecast Probability) (14)
15): Informedness or Bookmaker Informedness (BM) =BM=TPR+TNR-1 (15)
16): Markedness (MK): MK=PPV+NPV-1 (16)
17): ROC: It's the graph between true positive rate (Sensitivity) and false positive rate (Specificity).

3.3.1 Regression Metrics
Consists of statistical parameters like Mean Absolute Error, Mean Squared Error, Root Mean Squared Error etc. For making best out of the classifiers, simulation error is also measured in this study by measuring the worthiness of the classifier.
Kappa statistic (KS): is a chance corrected amount of agreeing in between the true classes and classifications.
Mean Absolute Error (MAE): nearness of forecasted predictions to the actual outcomes. There are also Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE).

IV. METHODOLOGY

This section rounds about the data mining algorithms used for the evaluation of the work. A brief touch is given to the actual working of the respective algorithm

4.1 Naive Bayes It is a special algorithm whose background lies in the famous foundation laid down by Bayes theorem and belongs to probabilistic method of classifiers. Bayes Classifier also known as generative model has its secret in computing class conditional probability in terms of posterior and prior probabilities [42].

4.2 J48 It is Decision Tree algorithm that uses a split criteria for splitting the data into the corresponding labels. The splitting condition can be single attribute known as Univariate or multiple featured known as Multivariate. Decision trees are one of the most popular classification approaches in machine learning [43]. The decision tree consists of a \root", "leaves", and internal nodes [44]. The internal nodes use certain features to split the instance space into two or more subspaces. Each leaf represents one class. The leaf may represent the most appropriate target value or indicate the probability of the target having a specific value. Fig. 4 is an example of the decision tree model. Decision trees are capable of handling datasets that may have missing values and errors, however, this method may overt fit training data and add unnecessary features. In radiological image analysis, decision trees are usually ensemble to form random forests for prediction and classification.

4.3 SVM Support vector machines (SVM) are kernel-based supervised learning techniques widely used for classification and regression [45,46]. The basic idea of SVM is to find an optimal hyper plane for linear separable patterns. . It attempts to maximize the geometric margin on the training set and minimize the training error. Then, a kernel function maps the original data into a new space for non-linearly separable cases, resulting in a two class classification problem $x_i, i = 1, 2, ..., N$ are feature vectors of the training set X , and of corresponding class indicator $y \in \{-1, +1\}$. The goal of SVM is to construct a classifier in the form of:

$$y(x) = sign[\sum_{i=1}^{N_a} \lambda_i y_i K(xi, x) + \omega_0] \quad (17)$$

The function $K(xi, x)$ is called the kernel function, and their different mathematical properties enable many pattern recognition and regression models. SVM with a linear kernel equation is computationally faster than SVM with quadratic kernel functions. SVM models using fewer but more significant features are most likely robust and less prone to over fitting [47].

4.4 Bagging Ensemble learning combines multiple classifiers and applies voting algorithms to achieve a final classification. Popular ensemble approaches include boosting and bagging [48]. Fig. 5 shows the basic idea of ensemble learning. In boosting, extra weight is assigned to incorrectly predicted points, and a set of weak classifiers are applied to deal with data in the training phase; the outputs of weak classifiers and the weighted inputs help calculate the final prediction. In bagging, the sub-classifier is independently constructed using a bootstrap sample of the data set and a majority voting method is applied

for the final prediction. Random forests are an ensemble learning method that consists of a multitude of decision trees. In standard tree construction, the node is split using the best split among all features. In a random forest, a random subset of features split each node. The random forest is one of the most powerful machine learning predictors used in detection, classification, and segmentation [49], particularly for brain [50, 51] and heart images.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

This paper carries a rigorous and in depth evaluation of ROC (receiver operating characteristics) of top most four data mining algorithms techniques of Naive Bayes, J48, SVM (Support Vector Machine) and an Ensemble techniques in the form of Bagging. All four algorithms are evaluated on two benchmark datasets of breast cancer and diabetes and data processing is accordingly done in three different data most used mining tools like Weka, Tanagra and MATLAB. The ROC in MATLAB always leaded the team in breast cancer case with an ROC value of 99.18 in all the four data mining algorithms case. This was followed by Bagging in Weka with an ROC value of 98.72, followed by Naïve Bayes in Weka with ROC value of 97.54. Tanagra got the 96.18 as the highest ROC in Ensemble case. Both Weka and Tanagra got lesser values than the MATALB in all the four cases. On the other hand, in case of diabetes dataset, the trend changed where it is Naïve Bayes with a value of 81.86 which topped the list in Weka, Tanagre and MATLAB. Similarly, in Tanagra case J48 with a value of 79.17 got the credit of leader. Again in SVM, MATLAB topped with a value 83.84and finally in Ensemble, MATLAB with a value of 99.18 became the leader.

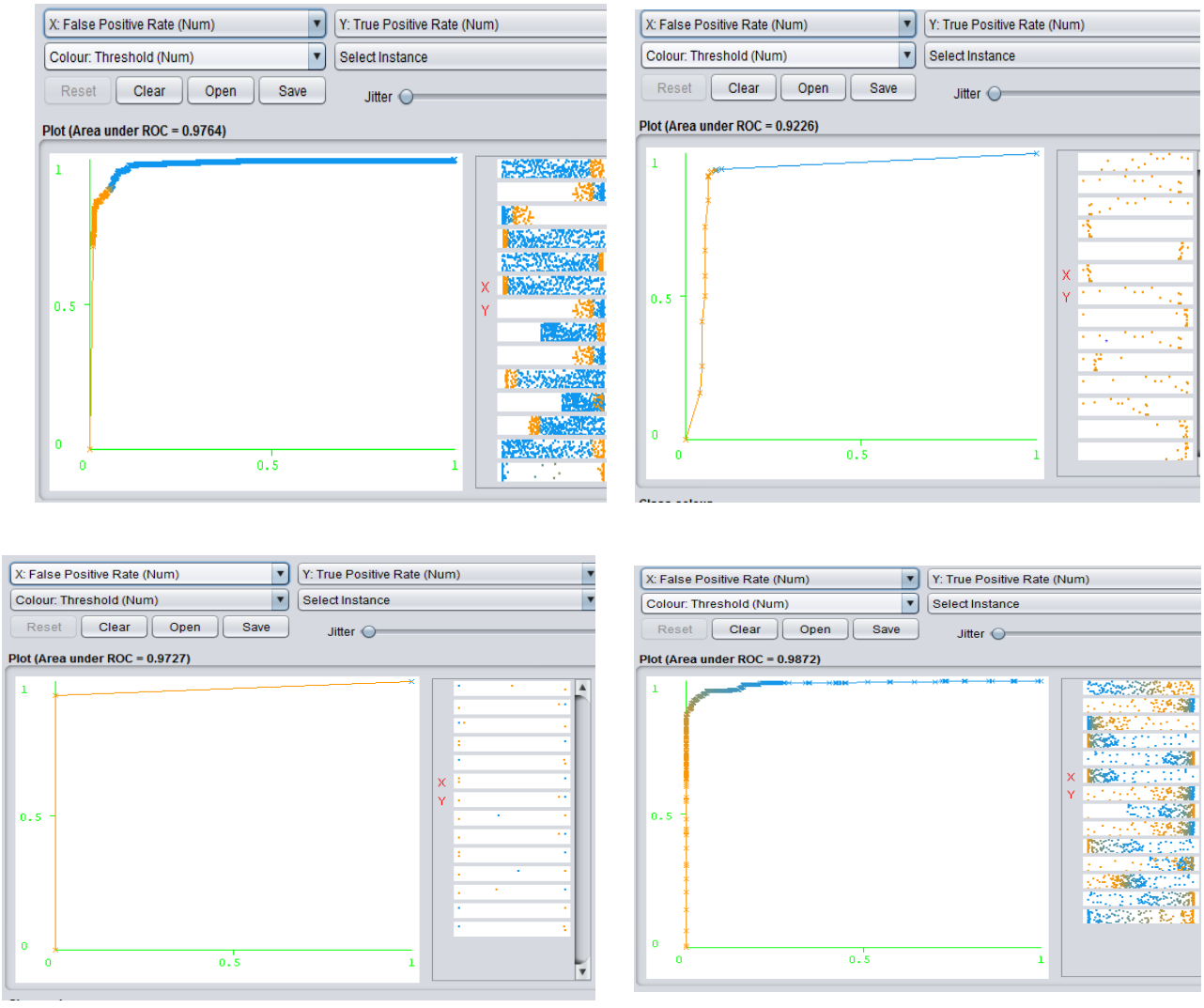
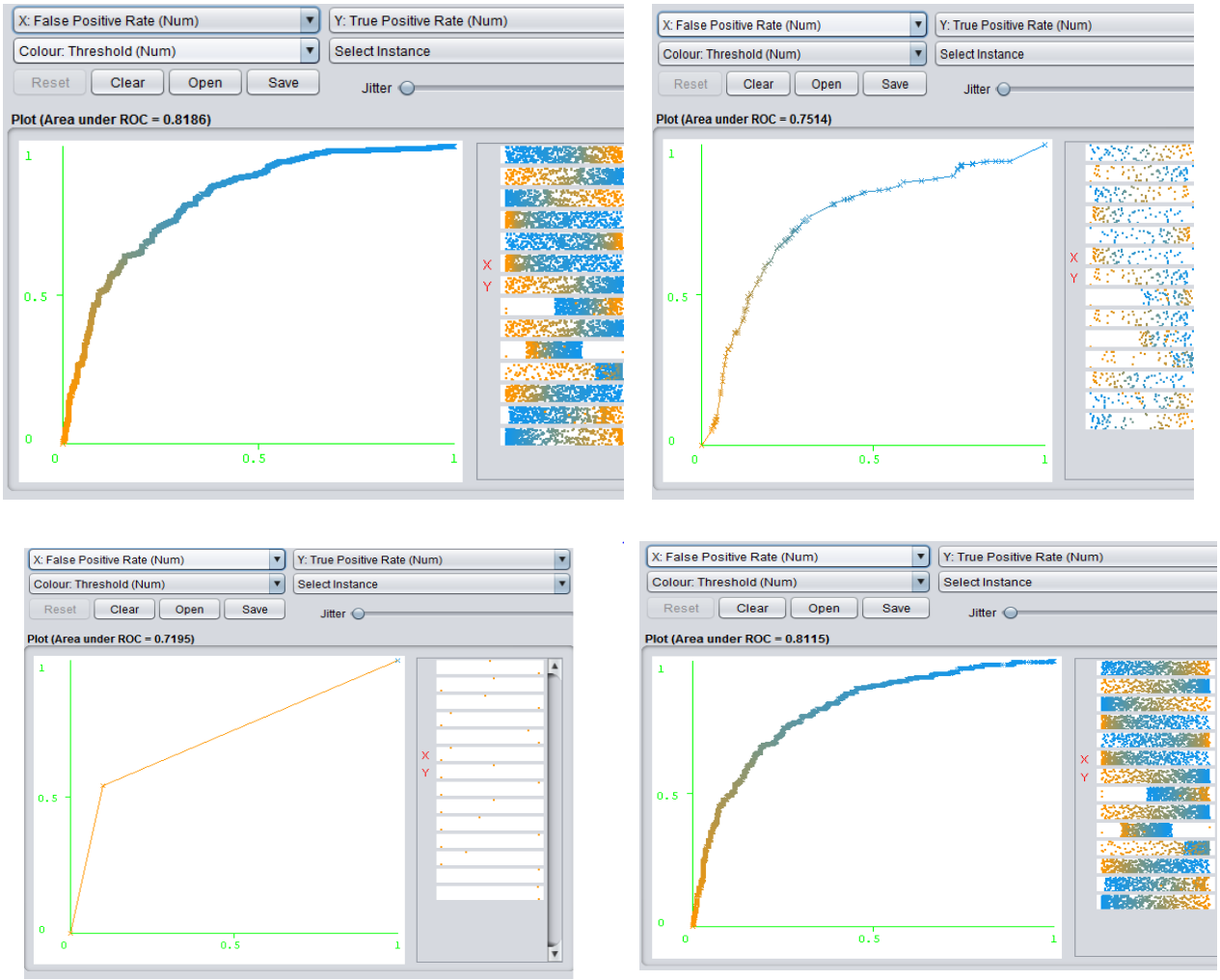


Fig. 2: ROC of (a) Naïve Bayes (b) j48 (c) SVM (d) Bagging, on breast cancer dataset for in Weka



breast cancer dataset for in Weka

Fig. 3: ROC of (a) Naïve Bayes (b) j48 (c) SVM (d) Bagging, on diabetes dataset for in Weka

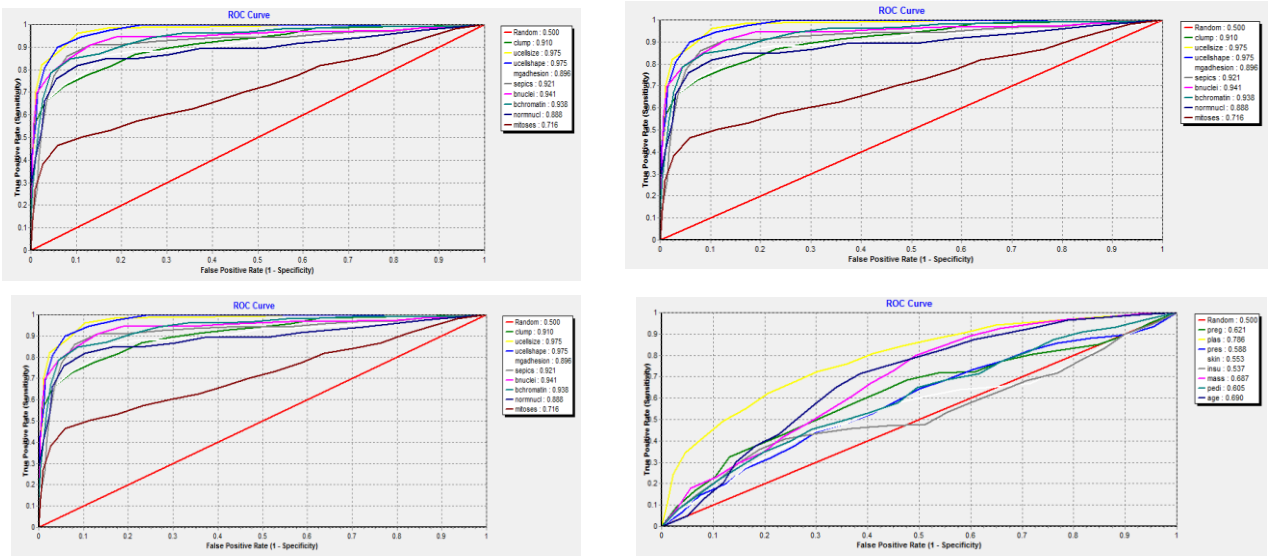


Fig. 4: ROC of (a) Naïve Bayes (b) j48 (c) SVM (d) Bagging, on breast cancer dataset for in Tanagra

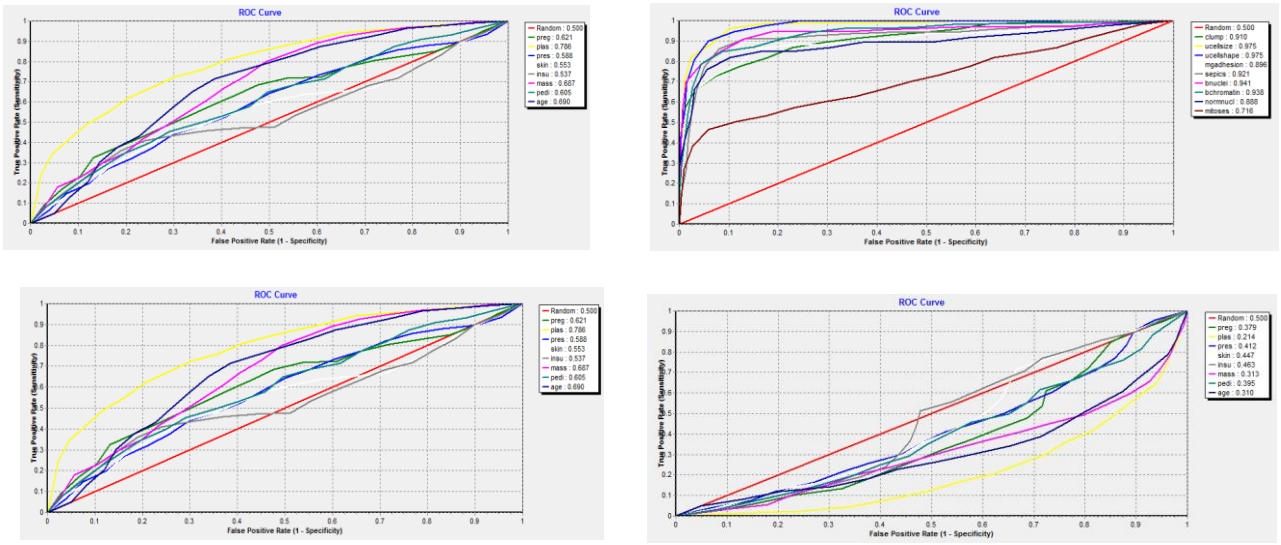


Fig. 5: ROC of (a) Naïve Bayes (b) j48 (c) SVM (d) Bagging, on diabetes dataset for in Tanagra

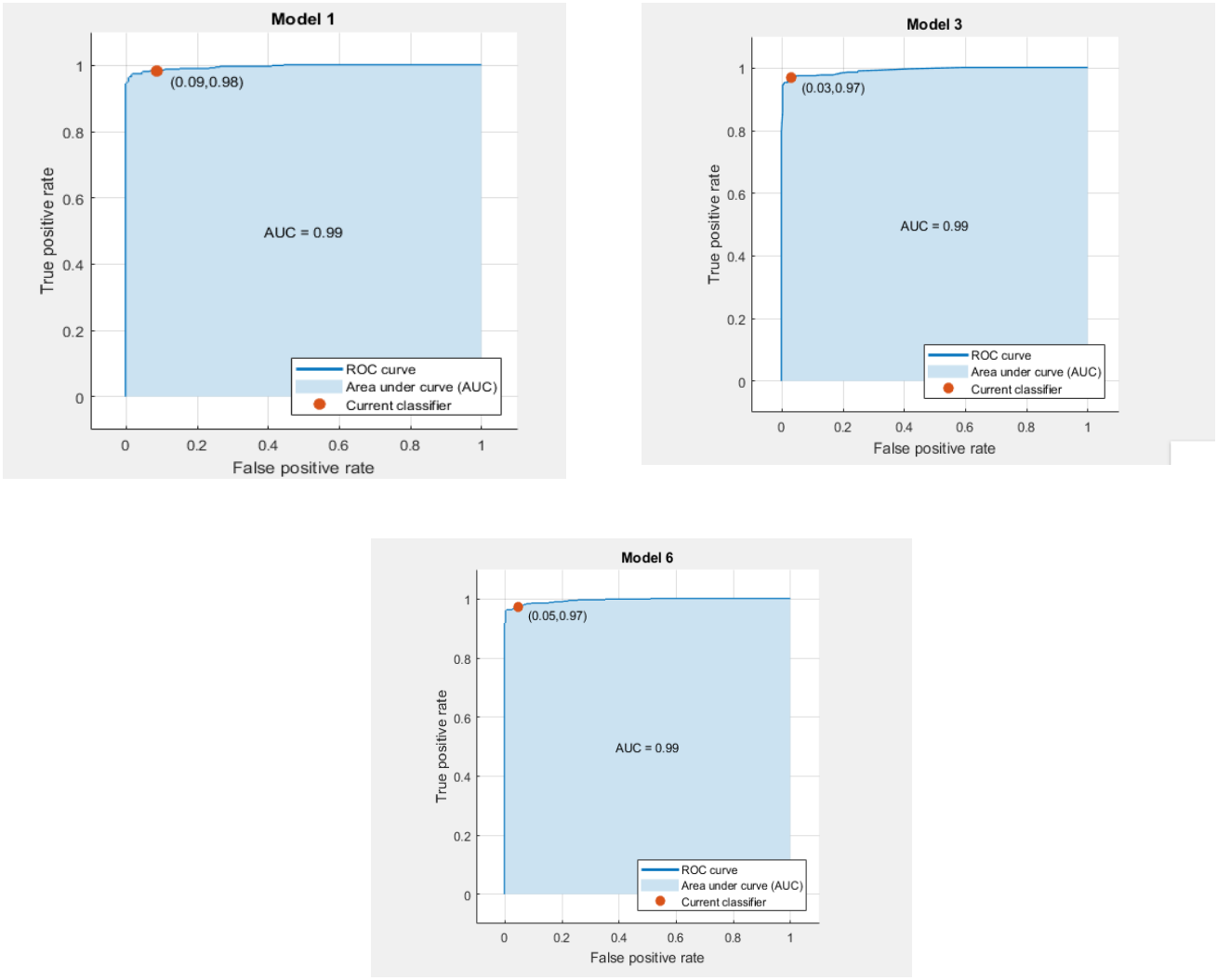


Fig. 6: ROC of (a) Naïve Bayes (b) j48 (c) SVM, on breast cancer dataset for in MATLAB

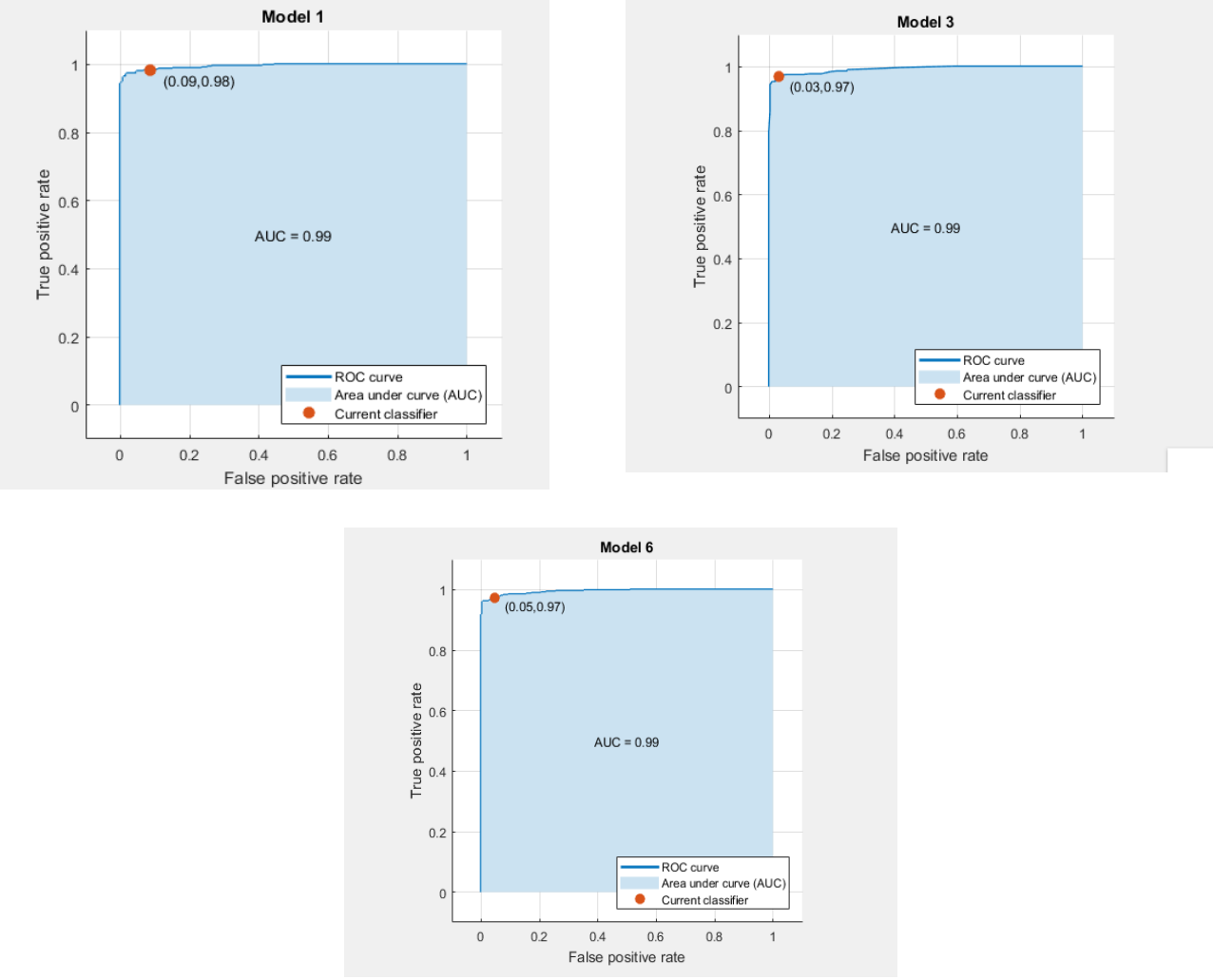
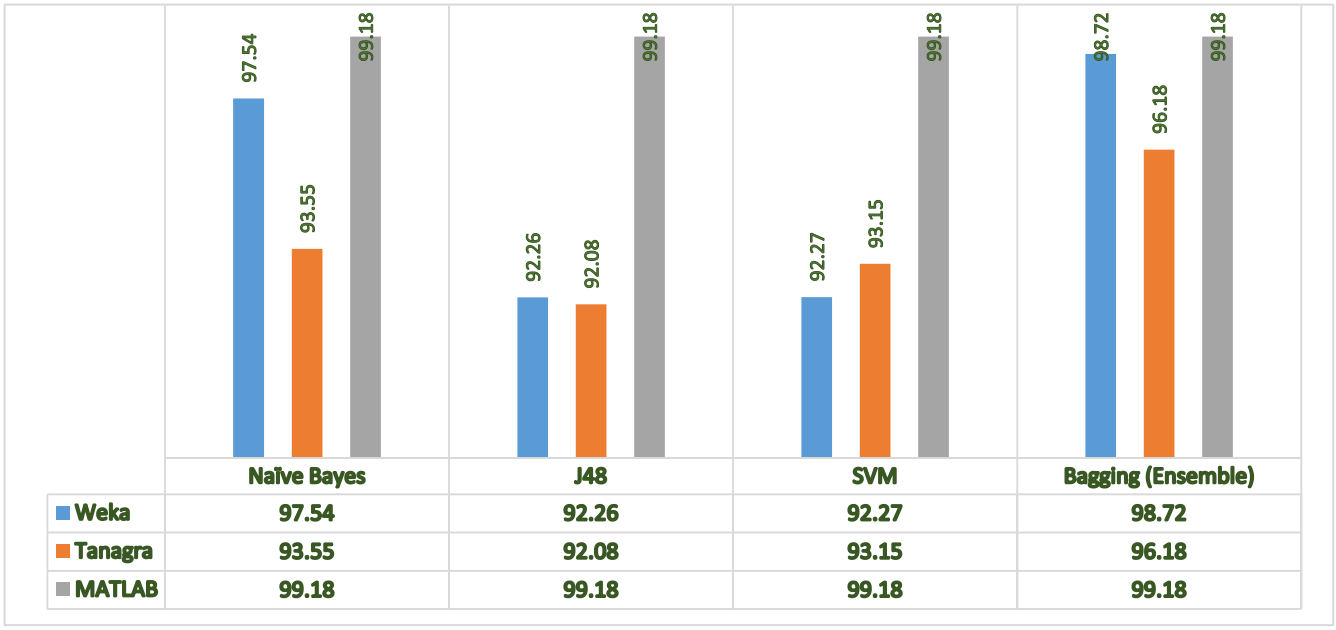


Fig. 6: ROC of (a) Naïve Bayes (b) j48 and (c) Bagging, on diabetes dataset for in MATLAB



	Naïve Bayes	J48	SVM	Bagging (Ensemble)
Weka	97.54	92.26	92.27	98.72
Tanagra	93.55	92.08	93.15	96.18
MATLAB	99.18	99.18	99.18	99.18

Fig. 7: ROC comparison of (a) Naive Bayes (b)j48 (c) SVM Poly k (d) Bagging on breast cancer dataset for positive case on Weka, Tanagra and MATLAB platforms

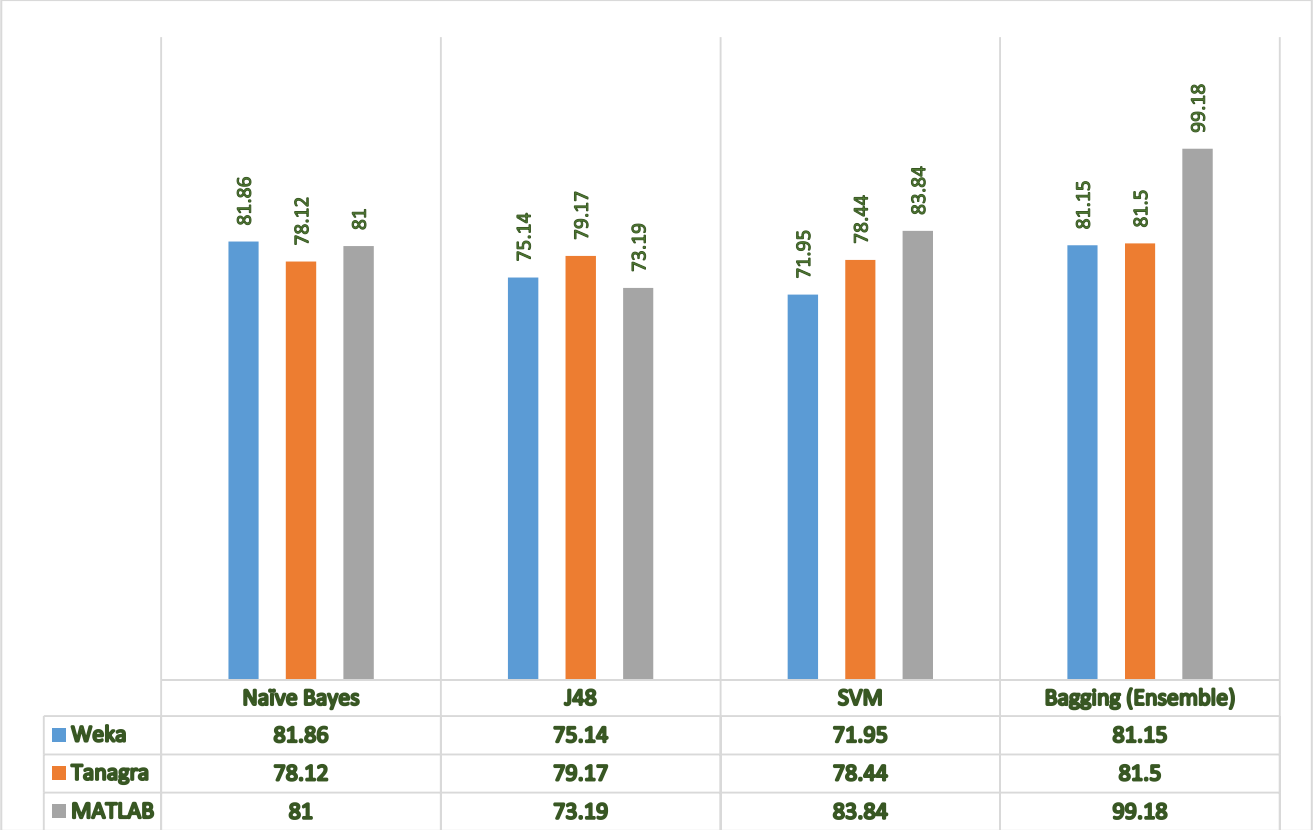


Fig. 8: Roc comparison of (a) Naive Bayes (b)j48 (c) SVM Poly k (d) Bagging on diabetes dataset for positive case on Weka, Tanagra and MATLAB platforms

The individual data mining algorithms when compared on all three platforms of Weka, Tanagra and MATLAB produced different outputs on each case, as mentioned in the above figures of 7 and 8. The Naive Bayes in case of Breast cancer dataset offered an Area under Curve (AUC) of 97.54, 93.55, and 99.18, in the case of underlying platforms of Weka, Tanagra and MATLAB respectively. Similarly, the decision tree algorithm J48 offered an Area under curve/ ROC curve of 92.26, 92.08, and 99.18, in case of data mining tools respectively as Weka, Tanagra and MATLAB. Similarly, the Support Vector machine learning algorithm when evaluated on breast cancer dataset on Weka, Tanagra and MTLAB platforms, yielded an AUC of 92.27, 93.15 and 99.18 respectively. Same way, after integrating the power of more data mining algorithms and make an ensemble in the form of bagging, yielded results of 98.72, 96.18, and 99.18 on breast cancer dataset on the same three data mining platforms. Overall, on breast cancer dataset it was is the MATLAB platform which showed the highest AUC rate in all data mining algorithms case (Fig. 7 and 8). Weka too got highest AUC rate in case of Naïve Bayes and Bagging techniques but J48 on Weka showed equal AUC as the J48 in that of Tanagra. The SVM in Weka case got drastic, as it produced output even worse than the SVM in Tanagra case (Fig. 7 and 8).

Similarly, when all four techniques were evaluated on three different data mining tools in case of diabetes dataset, the results also varied as in the breast cancer dataset case. In case of Naive Bayes, the Area under curve in diabetes dataset case came out to be 81.86, 78.12, and 81 after evaluating the same on three data mining tools of Weka, Tanagra and MATLAB. Similarly, J48 yielded an AUC of 75.14, 79.17, 73.17 in Weka, Tanagra and MATLA case. Moving ahead in same dimension, the hyperbolic SVM got an AUC values of 71. 95, 78.44, 83.84, on diabetes data set on all the three respective platforms. Finally, the Bagging ensemble on diabetes dataset offered an AUC of 81.15, 81.5, 99.18, in case of Weka, Tanagra and MATLAB platforms. Overall, on diabetes dataset, MATLAB became the winner of the group in the case of SVM and Bagging, while in case of Naive Bayes and J48, its AUC/ROC curve values are analogous and lies in the close vicinity of that of Weka and Tanagra. The Tanagra on diabetes case, overtook the Weka in case of SVM and J48 and also showed competitive performance in the Bagging and Naïve Bayes case. The algorithms on Weka platform showed random results (Fig. 7 and 8). The different machine learning algorithms on diverse data formats, performed in different ways. The best technique suitable for sorting out the problem at hand, lies in better understanding of the data one is working upon. The data compatibility with the machine learning algorithms makes the optimal results as output.

VI. CONCLUSIONS

The work here presented the evaluation of four most used data mining algorithms on two benchmark datasets. The breast cancer and diabetes datasets are selected as the two benchmark datasets and the evaluation is carried on three different used data mining tools of Weka, Tanagra and MATLAB. The Area under Curve (AUC)/ ROC Curve parameter is used to note and compare the performance of these selected machine learning algorithms. The results showed the variation on the performance of data mining algorithms which mainly depend on the dataset types, format and the machine learning tool used.
In future, more such rigorous experiments can be performed with a combination of new machine learning techniques. New tools on new benchmark datasets can be evaluated and visualized and the ensemble techniques of the best individual performer algorithms can be planted accordingly.

RERNENCES

1. R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2018, Ca-a Cancer J. Clin. 68 (1) (2018) 7–30.
2. American Cancer Society. (2016). *Cancer facts and figures 2016.* Atlanta: American Cancer Society.
3. Fan, L., Strasserweippl, K., Li, J. J., et al. (2014). Breast cancer in China. *The Lancet Oncology*, *15*(7), e279–e289.
4. Sizilio, G. R., Leite, C. R., Guerreiro, A. M., & Neto, A. D. D. (2012). Fuzzy method for prediagnosis of breast cancer from the fine needle aspirate analysis. *Biomedical Engineering*, *11*(1), 83.
5. E. Aliˇckovi´c, A. Subasi, Breast cancer diagnosis using ga feature selection and rotation forest, Neural Computing and Applications 28 (4) (2017) 753–763.
6. D. Shen, G. Wu, H.-Il Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
7. W.Q. Chen, et al., Cancer statistics in China, 2015, Ca-a Cancer J. Clin. 66 (2) (2016) 115–132.
8. M.T. Mandelson, et al., Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers, J. Natl. Cancer Inst. 92 (13) (2000) 1081–1087.
9. J.A. Harvey, V.E. Bovbjerg, Quantitative assessment of mammographic breast density: relationship with breast cancer risk, Radiology 230 (1) (2004) 29–41.
10. R.L. Siegel, K.D. Miller, S.A. Fedewa, D.J. Ahnen, R.G. Meester, A. Barzi, A. Jemal, Colorectal cancer statistics, 2017, CA Cancer J. Clin. 67 (3) (2017) 177–193.
11. Early detection: a long road ahead, Nature Review-Editorial. 18 (2018) 401.
12. Y. Zheng, Breast cancer detection with Gabor features from digital mammograms, Algorithms 3 (1) (2010) 44–62.
13. M. Tan, B. Zheng, J.K. Leader, D. Gur, Association between changes in mammographic image features and risk for near-term breast cancer development, IEEE Trans. Med. Imaging 35 (7) (2016) 1719–1728.
14. Karabatak, M. (2015). A new classifier for breast cancer detection based on Naive Bayesian. *Measurement*, *72*, 32–36.
15. Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications*, *42*(10), 4611–4620.
16. Chen, H., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014–9022.
17. Ahn, H., & Kim, K. (2009). Global optimization of case-based reasoning for breast cytology diagnosis. *Expert Systems with Applications*, *36*(1), 724–734.
18. Peng, L., Chen, W., Zhou, W., Li, F., Yang, J., & Zhang, J. (2016). An immune-inspired semi-supervised algorithm for breast cancer diagnosis. *Computer Methods and Programs in Biomedicine*, *134*, 259–265.
19. Na Liua,b, Er-Shi Qia, Man Xuc,□, Bo Gaod, Gui-Qiu Liue, 'A novel intelligent classification model for breast cancer diagnosis", Information Processing and Management 56 (2019) 609–623, Elsevoir.
20. M. Veta, J.P. Pluim, P.J van Diest, M.A Viergever, Breast cancer histopathology image analysis: A review, Biomedical Engineering, in: IEEE Transactions on. 2014 May, 61(5):1400-11.
21. M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, R. Monczak, Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images, in: Computers in Biology and Medicine. 2013, 43 (10):1563–1572.
22. B. Zhang, Breast cancer diagnosis from biopsy images by serial fusion of Random Subspace ensembles, in: 4th International Conference on Biomedical Engineering and Informatics (BMEI). vol. 1. Shanghai: IEEE, 2011. p.180–186.
23. Shallu∗, Rajesh Mehra, "Breast cancer histology images classification: Training from scratch or transfer learning?", ICT Express 4 (2018) 247–254, Elsevoir.
24. D.Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. (2016). ``Deep learning for identifying metastatic breast cancer." [Online]. Available: https://arxiv.org/abs/1606.05718
25. YUQIAN LI 1, JUNMIN WU1,2, AND QISONG WU3, 'Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning", VOLUME 7, 2019, pp: 21400- 21408, IEEE Acess.
26. Medical breast ultrasound image segmentation by machine learning Yuan Xua, Yuxin Wanga, Jie Yuana,□, Qian Chengb, Xueding Wangb,c, Paul L. Carsonc, Ultrasonics 91 (2019) 1–9, Elsevoir.
27. T. Araújo *et al.*, ``Classi_cation of breast cancer histology images using convolutional neural networks," *PLoS ONE*, vol. 12, no. 6, p. e0177544, 2017.

28. N. Nayak, H. Chang, A. Borowsky, P. Spellman, B. Parvin, Classification of tumor histopathology via sparse feature learning, in: 2013 IEEE 10th International Symposium on Biomedical Imaging, San Francisco, CA, 2013, pp. 410-413.
29. D.C. Ciresan, A. Giusti, L.M. Gambardella, J. Schmidhuber, Mitosis detection in breast cancer histology images with deep neural networks, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, 8150 LNCS(PART 2):411–418.
30. A Dynamic Gradient Boosting Machine Using Genetic Optimizer for Practical Breast Cancer Prognosis Hongya Lu, Haifeng Wang, Sang Won Yoon, *Expert Systems With Applications*, 2018, Elsevoir.
31. N. Karianakis, T.J. Fuchs, S. Soatto, Boosting convolutional features for robust object proposals, in: Tech. Rep. arXiv:1503.06350, University of California Los Angeles, Mar. 2015.
32. Y. Gao, W. Rong, Y. Shen, Z. Xiong, Convolutional neural net-work based sentiment analysis using adaboost combination, in: Proc IEEE IJCNN. Vancouver, Canada, pp. 1333–1338, Jul. 2016.
33. A new nested ensemble technique for automated diagnosis of breast cancer Moloud Abdara, Mariam Zomorodi-Moghadamb, Xujuan Zhouc,__, Raj Gururajanc, Xiaohui Taod, Prabal D Baruae, Rashmi Gururajanf, Pattern Recognition Letters, Elsevoir, pp:1-13.
34. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1993). UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/BreastCancerWisconsin(Diagnostic).
35. Tawseef Ayoub Shaikh and Rashid Ali, "Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk", *Proceedings of 2nd International Conference on Communication, Computing and Networking NITTR Chaandigarh, 3 March 2018*, Lecture Notes in Networks and Systems, Springer, Vol: 46, pp: 598- 589.
36. Tawseef Ayoub Shaikh, Rashid Ali and Aadil Rashid, "Soft Computing in Data Mining: A Tool for Computational Web Intelligence (CWI)", International Journal of Computer Applications (IJCA) Vol: 975, pp: 8887-8893.
37. BMF. Othman and Y. TMS, "Comparison of different classification techniques using WEKA for breast cancer", *3rd Kuala Lumpur international conference on biomedical engineering*, Springer, Berlin, Heidelberg, pp: 520–523, 2007.
38. M. Jordan J. Kleinberg B. Scho¨lkopf, "Pattern Recognition and Machine Learning", Springer.
39. M. Awawdeh, A. Fedi MATLAB-Based Graphical User Interface (GUI) for Data Mining as a Tool for Environment Management", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:8, No:1, 2014, pp: 133-140.
40. S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila, "A Review: Comparative Study of Diverse Collection of Data Mining Tools", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:8, No:6, 2014, pp: 1028-1033.
41. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information Processing and Management*, Elsevier, Vol- 45, pp: 427–437, 2009.
42. W.-Y. Loh, \Fifty years of classi_cation and regression trees," International Statistical Review, vol. 82, no. 3, pp. 329{348, 2014.
43. L. Rokach and O. Maimon, \Classi_cation Trees," in Data Mining and Knowledge Discovery Handbook, 2010, pp. 149{174.
44. N. Speybroeck, \Classi_cation and regression trees," International Journal of Public Health, vol. 57, no. 1, pp. 243{246, 2012.
45. C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
46. J. A. K. Suykens and J. Vandewalle, \Least squares support vector machine classi_ers," Neural Pro- cessing Letters, vol. 9, no. 3, pp. 293{300, 1999.
47. T. Torheim, E. Malinen, K. Kvaal, H. Lyng, U. G. Indahl, E. K. F. Andersen, and C. M. Futsaether, Classi_cation of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines," IEEE Transactions on Medical Imaging, vol. 33, no. 8, pp. 1648{1656, 2014.
48. E. Bauer, R. Kohavi, P. Chan, S. Stolfo, and D. Wolpert, \An empirical comparison of voting classi _cation algorithms: bagging, Boosting, and variants," Machine Learning, vol. 36, no. August, pp. 105{139, 1999.
49. TriHuynh, G. Yaozong, K. Jiayin, W. Li, Z. Pei, S. Dinggang, and Alzheimer's Disease Neuroimaging Initiative, \Multi-source information gain for random forest: an application to CT image prediction from MRI data," in International Workshop on Machine Learning in Medical Imaging, 2015, pp. 321{ 329.
50. D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and S. J. Price, \Decision forests for tissue-speci_c segmentation of high-grade gliomas in multi-channel MR," Medical Image Computing and Computer-Assisted Intervention, vol. 15, no. Pt 3,pp. 369{76, 2012.
51. E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, \Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images," NeuroImage, vol. 57, no. 2, pp. 378{390, 2011.