Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

Performance Analysis of Various Open Source Tools on Four Breast Cancer Datasets using Ensemble Classifiers Techniques

Ahmed Abd El-Hafeez Ibrahim[1], Atallah Ibrahin Hashad[2] and Negm Eldin Mohamed Shawky[3]
Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt

***Abstract*- The data mining is applied to discover the knowledge from information system. Classification is one of the tools which are used for data mining. Ensemble procedures verified to be superior to the single classification method for vast datasets. Hence, this paper presents an experimental study to investigate the quality of fusion methods for combining classifiers in an ensemble. Also, comparing between single classifiers and Ensemble Classifiers using majority voting with respect to accuracy in discovering breast cancer over four breast cancer datasets. We present a combination between classifiers to get the best subset of classifiers for each data set separately. By applying confusion matrix accuracy and 10-fold cross validation method. Also, we present a comparison among the three open source data mining tools named KNIME, ORANGE and TANAGRA. Analysis the performance of different classification algorithms shows that using Ensemble Classifiers Techniques improved the accuracy in three datasets out of four. Also, we prove that some open source tools are superior to others when using fusion.**

Keywords-Breast Cancer; Classification techniques; Fusion; Ensemble; UCI; Orange; Knime; Tanagra; majority voting.

I. INTRODUCTION

Breast cancer is a very serious malignant tumor originating from the breast cells. The disease occurs generally in women, but also men can rarely have it. During the prognosis of breast cancer, abnormal growth of cells in breast takes place and this growth can be in two types which are benign (non-cancerous) and malignant (cancerous). Nowadays, computer science and medical area are nested in order to offer a proper prognosis or diagnosis of the human diseases. Many computational methods are used for the identification of the health problems. Data mining has turned into a critical procedure for registering applications in the space region of medicine. In this study, it is aimed to identify the breast cancer with the help of data mining classification methods. The datasets named Wisconsin Diagnostic Breast Cancer Database (WDBC), Wisconsin Breast Cancer Database Original (WBC), Wisconsin prognosis Breast Cancer Database (WPBC) and Lubiana Breast Cancer Database University (LBCD) are obtained from university of California Irvine (UCI) respiratory and The Wisconsin Madison University [1, 2]. The classification techniques used over three open source data mining tools and also the ensemble of them. The three open source data mining tools named KNIME, ORANGE and TANAGRA. Furthermore, the feature selection algorithm is used to decrease the dimensionality of the datasets. In order to measure the performance, 10-fold cross validation technique is used on datasets. That is, the data are partitioned by the ratio 90:10% for training and testing. This is done ten times by a different 10% being tested each time.

The paper is prearranged as follows; in the following section, related works are indicated. And the next section describes Datasets and Evaluation Metrics. And the next section named proposed methodology presents the model development, the preprocessing steps. The results and the performance evaluations are discussed in the discussion and results section. Finally, the latter section introduces the conclusion of this study. Classification methods can achieve an early recognition of breast cancer with very high accuracy up to 98% in classifying malignant cases.

II. RELATED WORK

During the past few years, various contributions have been introduced in the literature regarding the Presentation of classification methods for breast cancer diagnosis. In this sector some of the related prior work on data mining procedures for breast cancer diagnosis is discussed.

A comparison between four of the open source data mining tools Weka, Tanagra, Orange, and Knime [3] over nine different datasets including WBC was introduced to judge the four tools applying six single classifiers. The study conducted that the type of dataset and the method the classification techniques were applied inside the toolkits affected the performance of the tools. The Weka has achieved the best results followed by Orange, Tanagra, and finally Knime, but we note that the comparison between tools in breast cancer dataset results only show that Tanagra is the best tool and Weka tool including the six classifiers where the rest of tools don't have the same number of chosen classifiers to get a reasonable comparison.

An analytical study among four diverse healthcare datasets including [WBC] data set over three various data mining tools Weka, Tanagra, Clementine and applying available single classifiers for each tool was introduced [4]. The study concluded that different classification techniques behave differently on different datasets depending on the nature of their attributes and size. The Tanagra toolkit has accomplished the best results for three datasets followed by Clementine, which achieved best results in a breast cancer dataset, but we note that the Clementine toolkit is a commercial tool and other tools are open source ones. Also the results ignore that Tanagra applying KNN classifier was better than Clementine in the breast cancer dataset too and has a lower error rate.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

A comparative study among three breast cancer datasets over diverse classifiers using Weka data mining tool and applying fusion between five classifiers was introduced [5]. In [WDBC] data set applying SMO as a single classifier only accomplished the best results and using fusion don't enhance the accuracy. In [WPBC] data set applying a fusion between MLP, IBK, SMO and J48 accomplished the best results and in [WBC] data set applying a fusion between J48 and MLP with the principle component analysis [PCA] is accomplished the best results.

A study proposed an approach for diagnosing the breast cancer from (WBCD) dataset using adaptive neuro-fuzzy inference system (ANFIS) which reached an accuracy of 98.25 % in tissue level [6].

A proposed model used between five classifiers on WBC dataset by removing the 16 instances with missing values from the dataset to form a different dataset with 683 cases instead of 699 was introduced using two data mining tools Weka and Tanagra [7] the classification technique, random tree outperforms has the highest accuracy rate.

A comparison between three classification techniques in Weka over WBC dataset was presented [8]. By eliminating the 16 cases with missing values from the dataset to form a diverse dataset with 683 cases instead of 699 cases. The (SMO) has the best accuracy.

A study expecting the Survivability of Breast Cancer cases applying Ensemble Method [9] proposed an ensemble classifier for expecting the survivability of breast cancer patient. Decision Tree Classifier, Naive Bayes and CMAR classifier are used to form the ensemble classifier on the base of voting strategy. Experimental results indicate that the suggested method accomplished better accuracy as compared to traditional classifier.

A discussion about knowledge extraction in medical data mining based on the reasoning for gynecological cancer [10], an expert diagnostic method was introduced collecting data over three diverse kinds of cancer. 626 instances collected are cervical cancer 290, for ovarian cancer 289 and for breast cancer 47 over four different classifiers using Weka software and got the highest accuracy with multiple layer perceptron (MLP) of the result (98%).

A study about Character Recognition using Ensemble Classifier proposed a model of classifier fusion for character recognition problem [11]. The work presented aimed to handle the disadvantages of classifiers and utilizes their strength with varying feature sets. The approach proposed the use of statistical procedures for the choice of the finest subgroup among diverse classification procedures and the subsequent combination of the decision of the techniques in this subgroup using weighted voting approach. The Experimental results showed that the performance of suggested ensemble classifier is superior as compared to other classifiers.

A Study of Classification Procedures presents a comparison among four classifiers LMT, FT, Simple Cart, Random Forest by applying them on three different datasets of carcinoma, breast cancer and cardiovascular disease [12]. The experimental results show that there's a major distinction within the accuracy of a same algorithm once applied on three completely different datasets. Also, it shows that the accuracy varies from associate procedure even on the same dataset. The results, proofs that FT algorithm is the best algorithm among the four algorithms. The results confined to the Weka tool only and the results might vary greatly once a similar dataset are classified on different tools. Also the study concluded that the accuracy of associate algorithm depends upon the number of attributes of the dataset.

An Analysis for Medical Diagnosis Using Open Source Tools presents a comparative study of diverse classification techniques using two open source data mining tools named Weka and Tanagra [13]. Among all the classification algorithms, the results are more accurate in Tanagra tool compared to Weka.

An Analysis of Effect of Reducing Dataset's Size on Classification Algorithms [14] proposes that some classifiers don't affected by size of dataset concentrates on the four algorithms, Naïve Bayes, Decision Tree , K Nearest Neighbor and Genetic Programming and the influence on their performance of time and accuracy when the number of cases are decreased paper will also explore the difference in result when working with binary class or multiclass datasets and propose the algorithms to follow when using certain type of dataset . The KNN shows to be the best and it is not dependent on the size of dataset. The Naïve Bayes dose depends on the size of the dataset, but works well than Decision Tree which also depends on the size of the dataset.

A study of diverse linear classification methods for breast cancer diagnosis applied logistic regression, multivariate linear regression [15], the K-nearest neighbor (KNN) method and discriminant analysis to diagnose tumor type using (WBC) database. In addition to linear approaches, quadratic discriminant analysis applied. Stepwise method for variable selection is used in regression method. The Results showed that: The accurate percentage of classification of testing data, depends on selected method and the number of independent variables and at least equal to 90.8% for logistic regression and a maximum 99.6% for KKN with K=9. Proofing that about 3.1% of doctors' diagnoses in the breast cancer may be incorrect. We note that dividing data into two parts; training (with 450 instances) and testing data (with 249 instances) achieving a high accuracy by selecting different variables.

A Survey on Ensemble Approaches for High Dimensional Data Classification in Biomedicine Field [16] surveys the different ensemble methods based on different feature selection criteria with same base classifier. The advantages and limitations of some ensemble methods is not uniform when it is applied to different datasets. The performance of these approaches is variable because of the characteristics of the features in the datasets and the approach of subset generation with a classifier.

A Comprehensive analysis of six open source data mining tools Weka, Knime, R, Keel, Orange and Rapid miner. The study describes the technical specification, features, and specialization for each certain tool along with its applications. By employing the study, the choice and selection of tools can be made easy of the six data mining packages that have been examined, KNIME is the package that would be recommended for people who are novices to such software to those who are highly skilled. The software is simply very robust with built-in features and with additional functionality that can be obtained from third-party libraries. Based on the analysis, Weka would be considered a very close second to KNIME

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

because of its many built-in features that require no programming or coding knowledge. In comparison, Rapid Miner and Orange would be considered appropriate for advanced users, particularly those in the hard sciences, because of the additional programming skills that are needed, and the limited visualization support that is provided. It can be concluded from above tables that though data mining is the basic concept to all tool yet, Rapid miner is the only tool which is independent of language limitation and has statistical and predictive analysis capabilities, so it can be simply used and implemented on any system, moreover it integrates maximum algorithms of other mentioned tools [17].

III. DATASETS AND EVALUATION METRICS

The breast cancer databases were gotten from the UCI machine-learning repository [1]

Dataset	Instances	Attributes	Attribute Type	Benign	Malignant	Missing Values
WBC	699	10	Integer	458	241	16
WDBC	569	32	Real	357	212	-

TABLE 1: Benign and Malignant Datasets

Dataset	Instances	Attributes	Attribute Type	Non Recurrence	Recurrence	Missing values
WPBC	198	34	Real	151	47	4
LBCD	286	9	Categorical	201	85	9

Table 2: Recurrence and Non recurrence Datasets

Evaluation method is based on the confusion matrix.
To calculate classifier performance. The accuracy term was used which is defined as the total sum of correct classified cases divided by the total sum of cases.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

IV. PROPOSED METHODOLOGY

We suggested a method for realizing breast cancer using four different data sets based on data mining as follows:

A. Data preprocessing

Preprocessing steps are applied to the data before classification:

1) Data Cleaning: removing or decreasing noise and the handling of missing values. There are 16 instances in WBC and 4 instances in WPBC that contain a single missing attribute value, denoted by "?"And there are 9 instances in LBCD that have two missing values which substituted by the median value for that feature built on statistics.

2) Relevance Analysis: Statistical correlation analysis is used to discard the redundant features from further analysis.

The WBC, WPBC and WDBC have one irrelevant feature named ‘Sample code number’ which has no influence in the classification procedure; therefore, the feature is not considered.

3) Data Normalization: hurries up training time by initialing the training process to reach feature within the similar scale. The goal of normalization is to convert the feature values to a small-scale range.

B. Multi-Classification Approach

Fusion Intelligent systems or Hybrid intelligent systems, in which more than one machine learning procedures are combined in one new approach are often effective and can overcome the limits of single methods [18]. There are two main models in combining diverse classification procedures: Classification Selection and classification Fusion [19]. Classification Selection paradigms use a single model to predicate the new case. However, fusion classification merges two or more outputs of all models produce a single output.

The process of combining more than one classifier is called multi-classification approach. The purpose of multi-classification is based on the argument that no single classifier that suites all learning problems [19]. The process of integrating two or more classifiers enhanced the classification accuracy in some cases. However, there is nope only one combination that suits all datasets.

Classifier selection is one of the simplest approaches for combining learning algorithms or classifiers. The idea is to evaluate two or more classifiers on the training dataset and then make use of the best performed classifiers on the testing dataset. This method is simple, straight forward, no output combination, and executes well in comparing to more compound classifiers [20].

Classifier fusion is a group of classifiers whose single predictions are combined to classify new cases (highest average ranking, average probability, or voting). It has become one of the active zones of study in supervised learning that study new ways of constructing classifiers for more accurate outcome. Voting is the simplest technique for multi-classification in heterogeneous and homogeneous models. We used un-weighted voting, which all classifiers are treated equally with no priority over other classifiers. Therefore, each classifier outputs a class value and the class with the most votes is the final outcome of the multi-classifier.

This type of voting is called Plurality Voting; the majority voting implies that at least 50%+1 (the majority) of the votes should belong to the winning class.

In Weka the class for uniting classifiers is called Vote Different combinations of probability deductions for classification are available. We have proven that fusion using Weka is superior to a single classifier with the same four breast cancer datasets [21]. Tools like Knime, Orange and Tanagra has no voting strategy, hence we build a source code using MATLAB to evaluate the voting over each of them

MATLAB is a fourth generation language and interactive environment for arithmetical calculation, visualization, and programming. MATLAB is used to analyze data, develop algorithms, and generate models and applications. Therefore, it is users coming from many families in engineering, science, and economics. [22].

1) According to results of single classification task, multi-classifiers task starts using the classifier accomplished the highest accuracy with other classifiers guessing to improve accuracy.
2) Repeating the procedure till the last level of fusion, conferring to the total number of classifiers to get the best accuracy through all levels of fusion.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

C. The Proposed Approach

We propose our process as follows.

1. Import the Database.
2. Substitute missing values with the median value.
3. Normalize each variable of the data set, with the aim of getting the values from 0 to 1.
4. Generate a separate training set and testing set by haphazardly drawing out the data for training and for testing.
5. Select and parameterize the learning process
6. Perform the learning process
7. Analyze the performance of the model on the test set

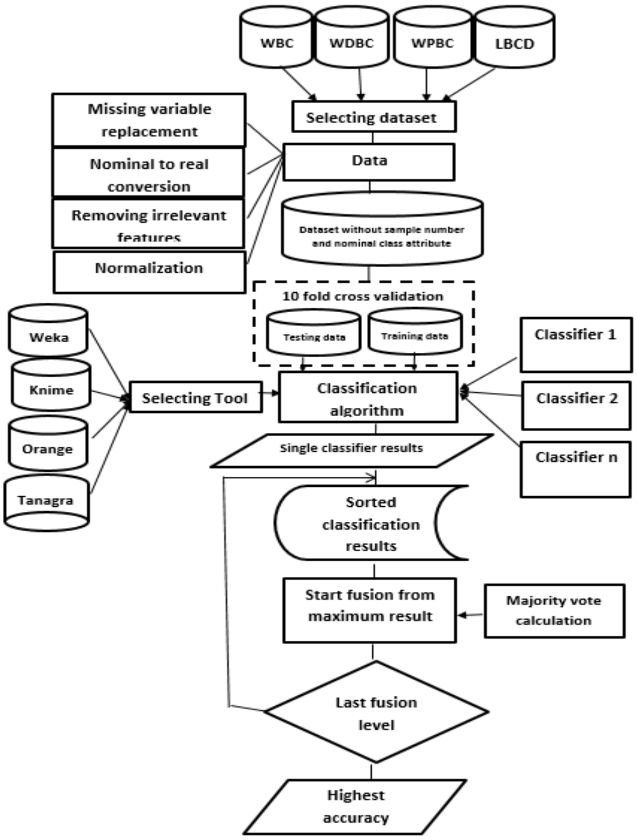

Figure 1.Proposed Breast Cancer diagnosis algorithm.

V. DISCUSSION AND RESULTS

Three experiments were performed, each experiment was repeated for each tool individually on four breast cancer datasets, and each experiment has three stages.

The first stage was performed on single classifier model to set a baseline of classification accuracy.

The second stage was performed on the fusion of classifiers with plurality voting.

The third stage a comparison between single and fusion classification for each tool.

Some classifiers have no highest accuracy for any dataset used, but we mentioned its results because it will have higher accuracy when combined with other classifiers in fusion case.

RESULTS OF ORANGE TOOLKIT

Single classifier

Seven classifiers are selected to conduct this experiment. Cross validation of 10 folds has been chosen as a test method, table 3 and shows the results of performing single classifiers on four datasets in regards to classification accuracy.

	NB	LR	KNN	SVM	CN2 RULES	RF	NN
WBC	97	96.28	96.14	96.42	95.14	94.71	96.71
WDBC	94.2	98.07	95.79	95.79	94.91	94.91	97.72
WPBC	65.58	78.74	73.18	80.34	75.24	77.84	77.79
LBCD	72.72	73.41	73.06	73.06	78.99	73.41	73.07

Table 3: single classifier accuracies percent for Orange toolkit

Table 6 shows the results of performing single classifiers on four datasets in regards to classification accuracy. It shows that Naïve Bayes performed best on WBC (97) while Logistic Regression has better on WDBC (98.07) and SVM has better on WPBC (80.34) and finally Cn2rules is the best on LBCD (78.99) respectively. We note that there is no single classifier has the best accuracy for more than one dataset.

- *Multi-classifier*

  In this experiment, the concept of voting technique is implemented by employing two classifiers instead of one classifier then three classifiers instead of two classifiers till the last combination of counting classifiers ends. The result of the highest classifier accuracy is superior to the other classifiers. Therefore, it will be selected to combine with the other classifiers respectively. We will consider that the combination between two classifiers called 2nd level of fusion and so on. Cross validation of 10 folds has been chosen as a test method. The Orange data mining tool doesn't implement voting technique that employee more than one classifier. Hence we apply our proposed source code in Matlab to evaluate voting plurality between classifiers. Table 4 shows the result of combining the highest classifier for each dataset and the other classifiers.

Fusion level	2nd	3rd	4th	5th	6th	7th	8th
WBC	96.57	97.14	96.85	97	96.57	96.71	96.28
WDBC	97.89	97.89	98.07	98.42	97.72	97.89	97.36
WPBC	81.31	80.3	81.82	81.31	81.82	82.32	81.82
LBCD	76.22	78.32	75.87	76.22	75.18	75.18	74.48

Table 4: multi-classifier highest accuracies percent for Orange toolkit

Using (WBC) dataset the accuracy of the fusion between the three classifiers NB, NN and KNN achieves accuracy (97. 14%).

Using (WDBC) dataset the accuracy of the fusion between the five classifiers SVM, NN, LR, CT and KNN similar to the fusion between SVM, NN, LR, CT and CN2 which achieves accuracy (98.42%).

Using (WPBC) dataset the accuracy of the fusion between the seven classifiers SVM, NB, LR, KNN, RF, CT and CN2 achieves accuracy (82.32 %).

Using (LBCD) dataset the accuracy of the fusion between the three classifiers CN2, NB and KNN achieves an accuracy (78.32%) which is lower than accuracy of a single classifier CN2 which has an accuracy (78.99) %.

We note that there is no unique level of fusion has the best results in all datasets, also KNN classifier is unique in enhancement accuracy in multi-classifiers fusion for all datasets and the combination between Naïve Bays and KNN is superior for three datasets (WBC, LBCD and WPBC).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

- Comparison

Table 5 shows the comparison of highest accuracies between single and multi-classifiers fusion task. The results indicate that the multi-classifier fusion task has achieved higher accuracy than single classifier accuracy in three breast cancer datasets out of four.

	WBC	WDBC	WPBC	LBCD
Single	97	98.07	80.34	78.99
Fusion	97.14	98.42	82.32	78.32

Table 5: single and multi-classifier highest accuracies percent for Orange toolkit

Results of KNIME Toolkit

- Single classifier

Six classifiers are selected to conduct this experiment. Cross validation of 10 folds have been chosen as test method, table 9 show the results of performing single classifiers on four datasets in regards to classification accuracy.

	NB	DTREE	PNN	FUZZY	MLP	SVM
WBC	94.6	93.4	97	96.3	95	96.9
WDBC	91.6	93.5	95.4	95.7	97.4	97.7
WPBC	75.3	70.7	75.8	72.3	71.2	76.8
LBCD	73.1	74.3	63.6	63.6	72.4	75.87

Table 6: single classifier accuracies percent for Knime toolkit.

Table 6 shows that PNN performed best on WBC (97) while SVM has better on three datasets WDBC (97.68) and on WPBC (76.8) and on LBCD (75.9) respectively.

- *Multi-classifier*

The Knime data mining tool doesn't implement voting technique that employee more than one classifier. Hence we apply our proposed source code in Matlab to evaluate voting plurality between classifiers. Cross validation of 10 folds has been chosen as a test method. Table 7 shows the result of combining the highest classifier for each dataset and the other classifiers.

Fusion level	2nd	3rd	4th	5th
WBC	97	97.12	97	96.71
WDBC	97.72	97.89	96.49	97.36
WPBC	77.27	78.79	77.78	77.78
LBCD	75.2	76.57	74.13	73.78

Table 7: multi-classifier highest accuracies percent for Knime toolkit.

Using (WBC) dataset the accuracy of the fusion between the three classifiers SVM, NB and PNN achieve accuracy (97.12%).

Using (WDBC) dataset the accuracy of the fusion between the three classifiers SVM, NB and MLP similar to the fusion between SVM, DT and MLP which achieve accuracy (97.89%).

Using (WPBC) dataset the accuracy of the fusion between the three classifiers SVM, NB and DT achieve accuracy (78.79%).

Using (LBCD) dataset the accuracy of the fusion between the three classifiers SVM, NB and MLP achieve accuracy (76.57%).

We note that the 3rd level of fusion has the highest accuracies for all datasets, we used, and also the combination between SVM and NB achieved the best fusion results in all datasets.

- *Comparison*

Table 8 shows the comparison of highest accuracies between single and multi-classifiers fusion task. The results indicate that the multi-classifier fusion task has achieved higher accuracy than single classifier accuracy in all breast cancer datasets.

	WBC	WDBC	WPBC	LBCD
Single	97	97.72	76.77	75.87
Fusion	97.12	97.89	78.79	76.57

Table 8: single and multi-classifier highest accuracies percent for Knime toolkit.

Results of TANAGRA Toolkit

- *Single classifier*

Thirteen classifiers are selected to conduct this experiment for three datasets and only seven classifiers applied to the fourth one. Cross validation of 10 folds has been chosen as a test method, Table 9 shows the results of performing single classifiers on four datasets in regards to classification accuracy. Table 12 shows that C-PLS (Partial least squares) performed best on WBC (97.25) while SVM has better on WDBC (97.68) and Logistic Regression has better on WPBC (82.11) and finally C4.5 is the best on LBCD (77.86) respectively.

Comparison between tools over datasets

Table 10 shows the comparison of highest accuracies between single classifiers task over the four datamining toolkits including Weka data mining tool from [21]. The results indicate that:

For single classifier stage

- The single classifier task has achieved higher accuracy than multi-classifier accuracy in one breast cancer datasets out of four.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

	C4.5	C-RT	CS-CRT	CS-MC4	ID3	KNN	RND-TREE	C-SVC	LINEAR-D-A	MULTINOMINAL-L-R	MLP	SVM	C-PLS
WBC	93.33	93.91	93.91	93.48	92.32	96.23	93.48	96.81	95.94	96.52	95.94	96.96	97.25
WDBC	91.96	91.79	91.79	91.25	87.5	97.32	92.86	97.5	95.18	7.14	97.32	97.68	96.61
WPBC	70.53	75.79	75.79	73.68	75.79	73.16	66.32	77.89	81.05	82.11	76.32	76.32	62.63
LBCD	77.86	68.93	68.93	70.36	71.79	71.43	75.71						

Table 9: single classifier accuracies percent for Tanagra toolkit

- Orange toolkit achieved the best accuracy in two datasets out of four, hence it is the superior tool we used.
- Weka toolkit achieved the highest accuracy in WBC using a Bayes Net classifier.
- Orange toolkit achieved the highest accuracy in WDBC using Logistic Regression classifier.
- Tanagra toolkit achieved the highest accuracy in WPBC using Multinomial Logistic Regression classifier.
- Orange toolkit achieved the highest accuracy in LBCD using the CN2 Rules classifier.
- SVM is the first classifier over all our experiments in single task which achieved six times from sixteen in the first rank.

Hence there is no single classifier is superior for more than breast cancer dataset.

For multi- classifiers stage

Table 11 shows the comparison of highest accuracies between multi-classifiers tasks over the four datamining toolkits including Weka data mining tool from [21]. The results indicate that:

- The multi-classifier fusion task has achieved higher accuracy than single classifier accuracy in three breast cancer datasets out of four.
- The orange toolkit achieved the best accuracy in two datasets and Weka toolkit achieved the best accuracy in the other two datasets we used.
- Weka toolkit has the highest accuracy in WBC using a combination between Bays Net, MLP, SMO and IBK classifiers in the fourth level of fusion.
- Orange toolkit has the highest accuracy in WDBC using a combination between LR, NN, SVM, CT and KNN classifiers which equal to the combination between LR, NN, SVM, CT and CN2 in the fifth level of fusion.
- Orange toolkit has the highest accuracy in WPBC using a combination between SVM, NB, LR, KNN, RF, CT and CN2 classifiers in the seventh level of fusion.
- Weka toolkit has the highest accuracy in LBCD using a combination between Bays Net, RF and J48 classifiers in the third level of fusion, but Orange toolkit has the highest accuracy with the same dataset in single classifier stage as we mentioned before.

Hence, there is no unique combination superior for all datasets or certain level of fusion could be superior too.

Tool	WBC	WDBC	WPBC	LBCD
Orange	NB	LR	SVM	CN2RULES
	97	98.07	80.34	78.99
Weka	BAYES NET	SMO	RF	J48
	97.28	97.72	78.29	76.63
Knime	PNN	SVM	SVM	SVM
	97	97.72	76.77	75.87
Tanagra	C-PLS	SVM	MULTINOMINAL-LR	C4.5
	97.25	97.68	82.11	77.86

Table 10: single classifier highest accuracies percent for four toolkit

Tool	WBC	WDBC	WPBC	LBCD
Orange	NB+NN+KNN	LR+NN+SVM+CT+CN2 LR+NN+SVM+CT+KNN	NB+SVM+LR+KNN+RF+CT+CN2	NB +cn2+ +KNN
	97.14	98.42	82.32	78.32
Weka	BN +SMO + +MLP+J48	BN +SMO	BN+RF+ ZeroR+MLP	BN +J48+RF
	97.57	97.72	80.84	78.67
Knime	SVM +PNN+NB	SVM+MLP+NB SVM+MLP+DT	SVM+NB+DT	SVM +PNN+NB
	97.12	97.89	78.79	76.57

Table 11: multi-classifier highest accuracies percent for three toolkit

VI. CONCLUSION

The experimental results illustrated that using Orange toolkit is superior to Weka, Knime and Tanagra toolkits which achieve the best results in three datasets out of four. Using multi-classifiers achieved higher accuracy than a single classifier in three datasets out of four. Each dataset has its own best single classifier which related to the data mining tool. SVM is superior in fusion with Knime tool. While The Bayes net is superior in fusion with Weka tool for all datasets. Also The Naïve Bayes is superior in fusion with Orange tool for three datasets out of four. We concluded that each tool has better results in fusion with a certain classifier combination. No multi-classifier fusion level or combination is superior for all datasets.

REFERENCES

[1] UCI Machine Learning Repository, Available at: http:// http://archive.ics.uci.edu/ml/, (Accessed 22 january 2015).

Mangasarian, O. L., Street, W. N., Wolberg, W. H. (1995) Breast cancer diagnosis and prognosis via linear Programming. Operations Research, 43(4), pp. 570-577.

[2] Wahbeh, Abdullah H., et al. "A comparison study between data mining tools over some classification methods." IJACSA) International Journal of Advanced Computer Science and Applications (2011): 18-26.

[3] Gupta, Shelly, Dharminder Kumar, and Anand Sharma. "Performance analysis of various data mining classification techniques on healthcare data."International Journal of Computer Science & Information Technology (IJCSIT)3.4 (2011).

[4] Salama, Gouda I., M. Abdelhalim, and Magdy Abd-elghany Zeid. "Breast cancer diagnosis on three different datasets using multi-classifiers." Breast Cancer (WDBC) 32.569 (2012):2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 03, March-2016

[5] Fatima, Bekaddour, and Chikh Mohammed Amine. "A neuro-fuzzy inference model for breast cancer recognition." International Journal of Computer Science & Information Technology 4.5 (2012): 163.

[6] Syed, S. S., S. Shanthi, and V. M. Chitra. "Application of Data Mining techniques to model breast cancer data." International Journal of Emerging Technology and Advanced Engineering 3.11 (2013): 362-9.

[7] Chaurasia, Vikas, and Saurabh Pal. "A Novel Approach for Breast Cancer Detection using Data Mining Techniques." International Journal ofr Innovative in Computer and Communication Engineering & 2.1 (2014).

[8] Rathore, Nisha, Divya Tomar, and Sankalp Agarwal. "Predicting the survivability of breast cancer patients using ensemble approach." Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on. IEEE, 2014.

[9] R. Vidya1 and G. M. Nasira2 (2015) "knowledge extraction in medical data mining: a case based reasoning for gynecological cancer - an expert diagnostic method "ARPN Journal of Engineering and Applied Sciences, Volume 10, issue 9, May, ISSN: 1819-6608, pp3997-4001.

[10] Tiwari, Sadhana. "Character Recognition using Ensemble classifier."

[11] D. Rajeswara Rao et al (2015)" Performance Analysis of Classification Algorithms

[12] VidyullathaPellakuri, DeepthiGurram, .D. Rajeswara, .M.R.Narasinga (2015) "Performance Analysis and Optimization of Supervised Learning Techniques for Medical Diagnosis Using Open Source Tools" International Journal of Computer Science and Information Technologies, Volume. 6, issue1, ISSN: 0975-9646, pp 380-383.

[13] Singhal, Neelam, and Mohd Ashraf. "Investigation of Effect of Reducing Dataset's Size on Classification Algorithms." Breast-cancer 699.10: 2.

[14] Zarei, Shaho, Mina Aminghafari, and Hakimeh Zali. "Volume: 4: Issue-2: February-2015 ISSN: 2278-0246 www. ijapbs. com APPLICATION AND COMPARISON OF DIFFERENT LINEAR CLASSIFICATION METHODS FOR BREAST CANCER DIAGNOSIS."

[15] Meshram, Shweta B., and Sharmila M. Shinde. "A Survey on Ensemble Methods for High Dimensional Data Classification in Biomedicine Field."International Journal of Computer Applications 111.11 (2015).

[16] Kalpana Rangra., and K. L. Bansal. "Comparative Study of Data Mining Tools."

[17] Dr. K. Ramar. Hybrid Intelligent System of Heterogeneous

Classifiers for Breast Cancer Diagnosis; International Journal of Computer Application (2250-1797) Volume 6–No.1, January-February 2016;pp30-35

[18] Grigorios Tsoumakas, Lefteris Angelis and Ioannis Vlahavas "Selective Fusion of Heterogeneous Classiers"

[19] S.B. Kotsiantis·I. D. Zaharakis·P.E. P i n t e l a s Machine learning: a review of classification and combining techniques Artif Intell Rev (2006) 26:159–190 DOI 10.1007/s10462-007-9052-3; Springer Science+Business Media B.V. 2007.

[20] Ahmed Abd El-Hafeez Ibrahim1, Atallah I. Hashad2, Nigm El-Deen M. Shawky3 and Aly Maher; "Robust Breast Cancer Diagnosis on Four Different Datasets Using Multi-Classifiers Fusion" International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181 Vol. 4 Issue 03, March-2015pp114-118

[21] http://www.iasbs.ac.ir/computer/tutorials/matlab/using_ml.pdf, MATLAB The Language of Technical Computing. Using MATLAB version 5.