ISSN No:-2456-2165

CRISP-MED-DM a Methodology of Diagnosing Breast Cancer

[1]Bouden Halima, [2]Noura Aknin, [3]Achraf Taghzati, [4]Siham Hadoudou, [5]Mouhamed Chrayah
[1]Information Technology and Systems Modeling Research Laboratory,
[1]Abdelmalik Essaadi University, Tetouan,Morocco

Abstract:- The aim of this study was to assess the applicability of knowledge discovery in database methodology, based upon DM techniques, to predict breast cancer. Following this methodology, we present a comparison between different classifiers or multi-classifiers fusion with respect to accuracy in discovering breast cancer for three different data sets, by using classification accuracy and confusion matrix based on a supplied test set method. We present an implementation among various classification techniques, which represent the most known algorithms in this field on three different datasets of breast cancer. To get the most suitable results we had referred to attribute selection, using GainRatioAttributeEval that measure how each feature contributes in decreasing the overall entropy.
The experimental results show that no classification technique is better than the other if used for all datasets, since the classification task is affected by the type of dataset. By using multi-classifiers fusion, the results show that accuracy improved, and feature selection methods did not have a strong influence on WDBC and WPBC datasets, but in WBC the selected attributes (Uniformity of Cell Size, Mitoses, Clump thickness, Bare Nuclei, Single Epithelial cell size, Marginal adhesion, Bland Chromatin and Class) improved the accuracy.

Keywords:- *Data Mining Methodology, CRISP-DM, Healthcare, Breast Cancer, Classification.*

I. INTRODUCTION

Numerous knowledge discovery focused data mining projects were established throughout the year as a result of the expansion of available data in the healthcare industry. Despite this, the medicine domain has a number of difficulties in its attempt to extract meaningful and implicit knowledge because of its particular qualities and intrinsic complexity, in addition to the absence of standards for data mining projects.

For this reason, we propose to apply in this article the Cross-Industry Standard Process for Data Mining (CRISPDM) approach to standardize data mining procedures in the healthcare industry. Widely used in many different industries, the CRISP-DM is a good foundational technique that may be improved upon to introduce domain specific standardizations.

The goal of this project is to use data mining algorithms to accelerate the breast cancer diagnosis process. The first result for the breast cancer test can be the outcome of the used data-mining model. Additionally, this can assist pathology laboratories and medical clinics in scheduling priority appointments and initiating treatment straight away for patients who test positive for cancer. According to studies, early detection programs can lower the disease's death rate [1],[2].

The literature reports a large body of research based on machine learning and data mining for the prediction of breast cancer from different datasets. In [3] an optimized KNN model is proposed for breast cancer prediction using a grid search approach to find the best hyper-parameter. The best accuracy obtained is 94.35%. In addition, Kaya and S., Yağanoğlu, M. [4] used six algorithms based on supervised machine learning (KNN, Random Forest, NB, Decision Trees, LR, and SVM) for classification, improving accuracy by combining linear discriminant analysis (LDA) and LR.

In [5] the authors used Particle Swarm Optimization (PSO) for feature selection in three classifiers (K-Nearest Neighbour (KNN), Naive Bayes (NB), and Fast Decision Tree (FDT) to optimise prediction performance on WPBC data and obtained the highest accuracy of 81.3% using the NB classifier.

Some authors work on the Wisconsin datasets, first authors presented a comparison between different classifiers (J48, MLP, BN, SMO, RF and IBK) on three different databases of Wisconsin, by using multi-classifiers fusion the results show that accuracy improved [6], other authors proposed an ensemble of neural networks comprised of the RBFN, GRNN and FFNN for breast cancer diagnosis [7]. The hybrid model proposed improves accuracy rate reasonably and sensitivity rate substantially on the common WDBC dataset. In [8], the authors used a deep learning algorithm with several activation functions such as (Tanh, Rectifier, Maxout and Exprectifier) to classify breast cancer. They achieved the highest classification accuracy (96.99%) using the Exprectifier function with (breast cancer Wisconsin dataset).

In [9], authors propose an ensemble method named stacking classifier. They implemented different classification methods over the WDBC dataset and fine-tuned their parameters to achieve a better classification rate. By integrating the findings of those classifiers, they got 97.20% accuracy.

The rest of the paper is structured as follows. Section 2 "Materials and methods" briefly explains classifiers techniques, Cross-industry standard process for data mining methodology, and Weka Software. Extension of CRISP-DM data mining methodology for breast cancer diagnosis is presented in section 3. Section 4 describes the experimental

study, numerical results, and different comparisons. Finally, Section 5 draws the conclusion.

II. MATERIALS AND METHODS

A. Classifiers Techniques

The Multilayer Perceptrons (MLPs) are supervised learning classifiers composed of an input layer, an output layer, and one or more hidden layers. The MLPs provide changeable weighting coefficients to input layer components and extract valuable information throughout the learning process. In pattern recognition, the most popular neural network technique is the feed-forward back-propagation network, or MLP [10, 11]. The output is produced by giving the motivation level over a transmission function the weighted sum of the inputs plus the bias term. Additionally, a layered feed-forward neural network (FFNN) [12] is used to arrange the units. Different neurons found in the input layer as the number of features in a feature vector. With regard to the second layer (hidden layer), it has h number of Perceptions, where the value of h is determined by trial. Finally for the output layer we have only one neuron representing either benign or malignant value (in case of diagnosis datasets). We used sigmoid activation function for hidden and output layers. Weights between various layers are updated using the batch learning method [13].

Instances are categorized using K-Nearest Neighbor (KNN) classification [14, 15], which takes similarity into account. It is among the most widely used pattern recognition algorithms. This type of lazy learning postpones all computation until classification and only approximates the function locally. The majority of its neighbors classify an object. K is a positive integer at all times. A set of objects for which the correct classification is known is used to choose the neighbors. This classifier is known as IBK in WEKA.

Decision tree J48 implements Quinlan's C4.5 algorithm [16] for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm used classification, either J48 create decision trees from a set of categorized training data using the theory of information entropy. For making a decision we can dividing the data into smaller subsets of each attribute (both continuous and discrete can be handle by J48).Training data with missing attribute values and attributes with differing costs. Further, it provides an option for pruning trees after creation.

Random Forest [17] contains many decision trees and productions, it's a combined classifier. It introduces two bases of randomness: "Bagging" and "Random input vectors", respectively; a tree is grown by a bootstrap model of training data. At each node, greatest divided is selected from a random model of variability rather than all variables [13].

Support Vector Machine (SVM) [18] it is a very powerful method applied in a wide selection of applications with a hyper plane classifier, or linear separability, to achieve the latter; we need two basic ideas: margin maximization and kernels: mapping input space to a higher-dimension space (or feature space). Inputs data are projects by SVM into a kernel space; after that SVM builds a linear model in the same kernel. This model aims to break up the target classes with the widest possible margin. A backward the goal of the SVM model is to identify a continuous function around which the greatest number of data points fall inside an epsilon-wide tube. Different function approximators (regression) or decision boundaries (classification) can be produced by varying kernel types and kernel parameter selections. This classifier is known as SMO (Sequential Minimal Optimization) in WEKA [19], [20]. This new technique is simple and fast for training an SVM. By enhancing the least subset that contains two features at each iteration, the double quadratic optimization problem can be solved. It is easily and analytically implementable. Solving optimization problems involving a lot of quadratic programming is necessary for training a support vector machine.

Naive Bayes (NB) [21] (Based on the Bayes theorem) the classifier is a probabilistic classifier. The Naïve Bayes classifier generates probability estimates instead of predictions. They calculate the likelihood that a particular instance belongs to each class value. The benefit of the Naive Bayes classifier is that it only needs a small amount of training data to estimate the parameters required for classification. It makes the assumption that an attribute's impact on a particular class is unaffected by the values of the other attributes. Class conditional independence is the name given to this presumption [22].

B. Cross-industry standard process for data mining Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996 by Daimler-Chrysler, SPSS (then ISL), and NCR, pioneers in the emerging field of data mining. Designed as an industry- and tool-independent data management process model, it has become the industry-standard methodology for data management and predictive analytics in a variety of sectors. CRISP-DM makes large-scale data management projects faster, cheaper, more reliable and more manageable, while also benefiting small-scale data mining investigations. The founders aimed to create a standard, non-proprietary and freely accessible model for engineering data management applications. The current version includes the methodology, reference model and implementation user guide. The methodology defines phases, tasks, activities and deliverables outputs of these tasks [23].

As illustrated in following calligrapher CRISP-DM proposes an iterative process flow, with non-strictly defined loops between phases, and overall iterative cyclical nature of DM project itself. The result of each step determines which step, or which specific task within that step, must be followed. These are the six phases of CRISP-DM [24]:

ISSN No:-2456-2165

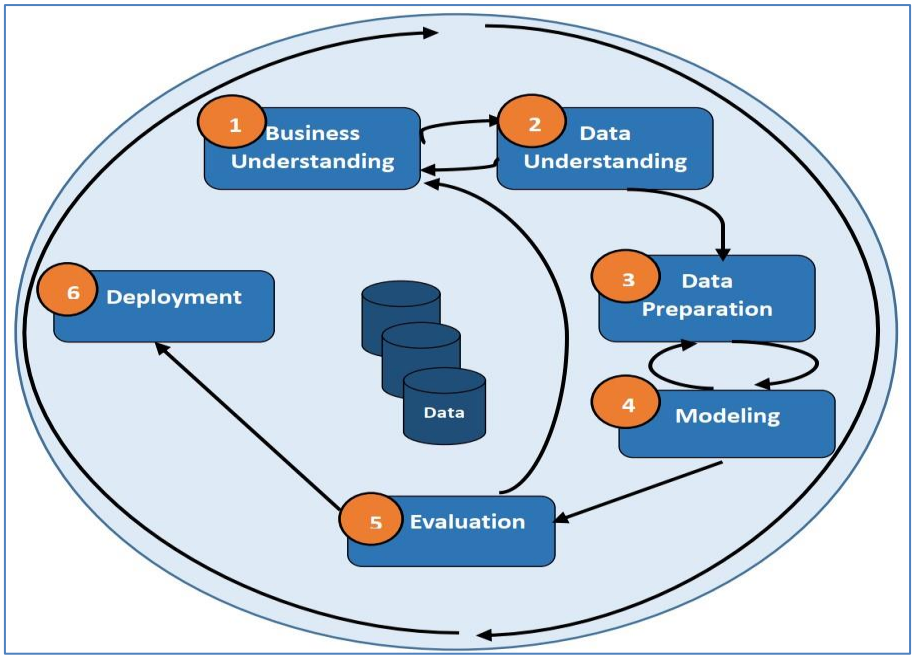

Fig. 1: The six phases of CRISP-DM

➢ *Business Understanding*

This initial phase sets the industrial objectives and success criteria, evaluates the resources, constraints and assumptions required to achieve the objectives, converts the industrial goals and criteria into a data mining problem definition, and outlines a plan for resolving the issue to meet the technical objectives.

➢ *Data Understanding*

The data understanding phase carries out the initial data collection, produces a description of the data, examines any hypotheses with visualizations, and checks the quality level of the data.

Business Understanding and Data Understanding are inextricably linked. The formulation of the data mining problem and the project plan both necessitate some knowledge of the available data.

➢ *Data Preparation*

The original raw data is rarely ready for use. This phase involves cleaning, transforming and enriching the data to make it suitable for modeling. The initial raw data's features and quality have a significant impact on the data preparation phase's operations.

➢ *Modeling*

Various modeling techniques are chosen and employed in this phase. The prepared data are used to train different models, which are then evaluated and optimized according to the defined performance criteria.

➢ *Evaluation*

The evaluation phase assesses whether the industrial objectives have been achieved, ensuring that the process has gone according to plan.New objectives may then be created on the basis of newly discovered models. This is in fact an iterative process, and the decision as to whether or not to take them into account must be taken at this stage before moving on to the final phase.

➢ *Deployment*

Creating the model is generally not the end of the project. Despite the cases where the objective of DM project was to learn more about the data available, the acquired knowledge should be structured and presented to the end user in an understandable form. Frequently, it will be the user and not the data analyst who executes the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

C. WEKA Software

WEKA an open source software, developed at the University of Waikato, New Zealand [25].Weka is a set of machine learning algorithms for data mining tasks. The algorithms can be either directly applied to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. What WEKA offers is summarized in the following diagram:

ISSN No:-2456-2165

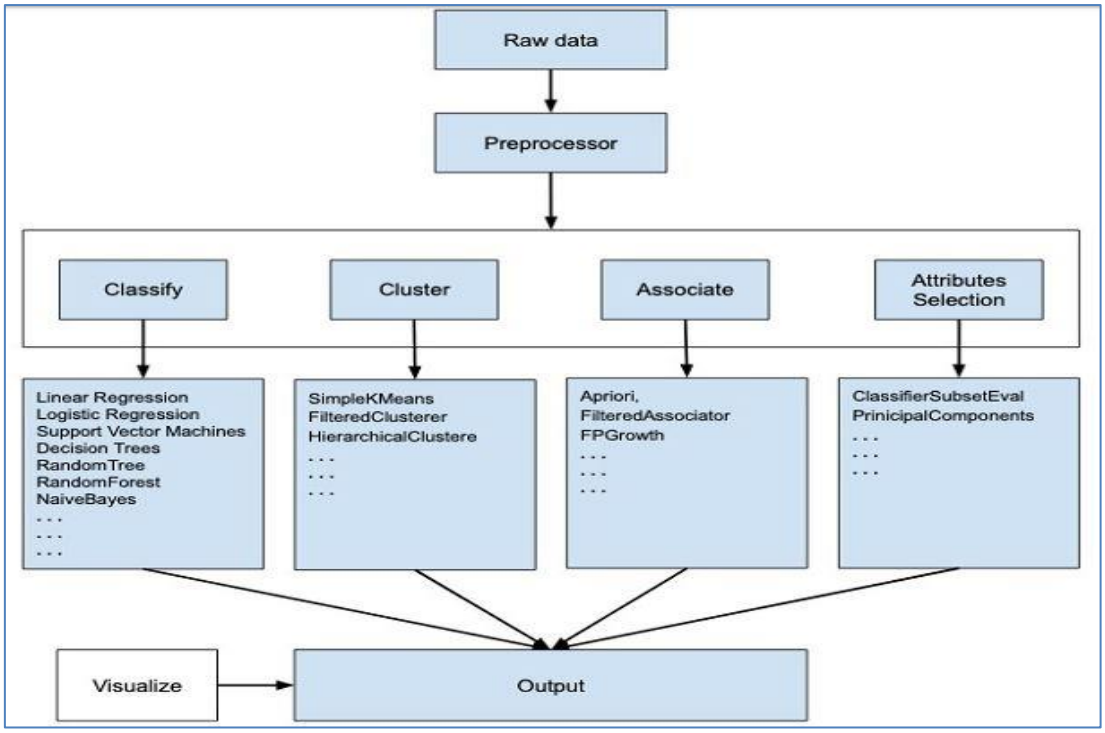

Fig. 2: WEKA Software

III. CRISP-DM DATA MINING METHODOLOGY EXTENSION FOR BREAST CANCER DIAGNOSIS

A. Phases 1. Project scope definition

The phases in CRISP-DM where the DM project is defined and conceptualized are "Business understanding" (phase 1) and "Data understanding" (phase 2).

The implementation phases, which make up the remaining stages, are designed to accomplish the goals established in the initial phases. The implementation stages are extremely incremental and iterative, just like in the original CRISP-DM. On the other hand, modifications to Phases 1 or 2 result in modifications to the project's goals and resources. To avoid giving the first phase an unclear meaning, "Business understanding" was renamed to "Problem understanding."

B. Phases 2. CRISP-DM for discovering breast cancer methodology

We proposed a method for discovering breast cancer using three different datasets based on data mining using WEKA. The Proposed Breast Cancer Diagnosis Model consists of CRISP-DM phases the add is in the modeling phase. We propose a fusion at classification level between these six classifiers [decision tree (J48), Multi-Layer Perception (MLP). Bayes net (BN), Sequential Minimal Optimization (SMO), Random forest (RF) and Instance Based for K-Nearest neighbor (IBK)] on three different databases of breast cancer (WBC), (WDBC) and (WPBC)to get the most suitable multi-classifier approach for each dataset.

➢ *Data Understanding*

We used tree datasets from UCI Machine Learning Repository [26]:

- Wisconsin Breast Cancer (WBC)
- Wisconsin Diagnosis Breast Cancer (WDBC)
- Wisconsin Prognosis Breast Cancer (WPBC)

A set of numerical features or attributes are associated with certain classification patterns or instances in each dataset. Table.1 provides a brief explanation of these datasets.

Table 1: Description of the breast cancer datasets

Dataset	No of instances	No of attributes	Missing values
WBC	699	11	16
WDBC	569	32	-
WPBC	198	34	4

➢ *Data Preparation*

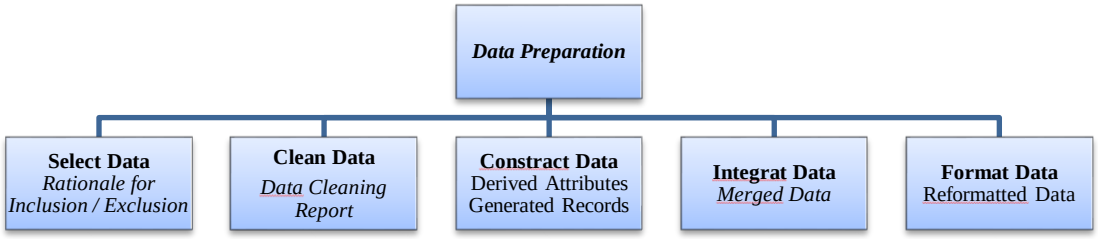

Fig. 3: Data Preparation

The five steps of this phase can be resumed into two principal steps: data Selection and preprocessing steps.

- **Select Data:** We Import the first Dataset "WBC" and we apply the six classifiers successively. According to results of single classification task, multi-classifiers fusion process starts using the classifier achieved best accuracy with other single classifiers predicting to improve accuracy. Repeating the same process till the latest level of fusion, according to the number of single classifiers to pick the highest accuracy through all processes.

This process will be repeated into the other datasets (WDBC) and (WPBC).

- **Preprocessing steps:** Preprocessing steps are applied to the data:

✓ Data Cleaning: Removing or reducing noise and the treatment of missing values. There are 16 WBC instances and 4 WPBC instances with a single missing attribute value, denoted by "?"

✓ Feature extraction and Relevance Analysis: Statistical correlation analysis is used to discard the redundant features from further analysis. Feature extraction considers the whole information content and maps the useful information content into a lower dimensional feature space. Feature selection is based on omitting those features from the available measurements, which do not contribute to class separability. That is, redundant and irrelevant features are ignored. (This step was applied in the modeling phase)

✓ Modeling: According to CRISP-DM, Modeling phase is iterative and recursively returns back to the data preparation phase. In addition, there is iteration within modeling phase between the tasks "Build Model" and "assess Model". However, the process flow of these iterations is not defined in the reference model and is not self-evident. Spečkauskienė and Lukoševičius [27] proposed iterative 11-step DM process model, tailored for finding optimum modeling algorithm. The authors suggested the following sequence:

- Collect and access a set of classification algorithms.
- Analyze the data set.
- Identify suitable algorithms for the dataset.
- Test the entire dataset using a chosen set of classification algorithms with standard parameter values.
- Choose the optimal algorithms for further analysis.
- Train the selected algorithms with a restricted dataset, eliminating uninformative attributes.
- Adjust the standard algorithm values using the optimal dataset assembled for each algorithm based on the most informative data identified in step 6.
- Evaluate the obtained results.
- Randomize the attribute values of the dataset.
- Repeat steps 6 and 7 with a new dataset.
- Assess and compare the performance and efficiency of the algorithms.

Our modeling steps are adjusted based on the 11 steps of [27] to add the steps of multi-classifier fusion as follows:

A series of classification algorithms was collected [decision tree (J48), Multi-Layer Perception (MLP) Naïve Bayes (NB), Bayes net (BN), Sequential Minimal Optimization (SMO), Random Forest (RF) and Instance Based for K-Nearest neighbor (IBK)…].

The datasets (WBC, WDBC and WPBC) were analyzed by using these algorithms with the standard parameter values. Algorithms appropriate for the dataset are shortlisted based on their accuracy and on the previous studies, that shown a good result with the same datasets (J48, MLP, BN, SMO, RF and IBK).

The best performing algorithms for other datasets are selected and optimized using a restricted dataset, excluding attributes that have been identified as uninformative. The multi-classifier fusion process starts with the best classifier

ISSN No:-2456-2165

from a single classification task, combining its predictions with other classifiers to improve accuracy. This step is repeated up to the last fusion level, adjusting the number of classifiers, to achieve the best overall accuracy.In the last step, performance and efficiency of the algorithms in each datasets was evaluated and compared.

- **Evaluation:** A Confusion Matrix is a summary of the results of predictions on a classification problem. It is used to show the relationships between outcomes and predicted classes.

The level of effectiveness of the classification model is calculated with the number of correct and incorrect classification in each possible value of the variable being classified in the confusion matrix [28].

In the context of our study, the entries in the confusion matrix have the following meaning :

- ✓ TP (True positive): Is the number of correct prediction in benign class.
- ✓ FP (False positive): Is the number of incorrect prediction in benign class.
- ✓ TN (True negative): Is the number of correct prediction in a malignant class.
- ✓ FN (False negative), is the number of incorrect prediction in malignant class.

Table 2: Confusion matrix

		Predicted	
		Benign	Malignant
Actual	Benign	TP	FN
	Malignant	FP	TN

The accuracy (AC): is the proportion of the total number of predictions that were correct. It is determined using equation (1), and the statistical parameters for measuring the factors that affect the performance (sensitivity and specificity) are presented in equation. (2) and (3), respectively.

Sensitivity is referred to the true positive rate, where specificity is the negative rate.

$$AC = \frac{TP + TN}{FP + FN + TP + TN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FP} \quad (2)$$

$$Specificity = \frac{TN}{TN + FN} \quad (3)$$

IV. EXPERIMENTAL AND COMPARATIVE RESULTS

There were two experiments executed and three datasets were used in each of the two tasks: the first for the single classification job and the second for the multi-classifier fusion task:

A. Experiment (1) using Wisconsin Breast Cancer (WBC) dataset

Figure.1. present the comparison between accuracies for the six classifiers (BN, MLP, J48, SMO, IBK and RF) in tow circumstance first without Attribute selection second with attribute selection, based on supplied test set as a test method.

- BN accuracy (97.28%) is higher than others classifiers (SMO,RF, IBK, MLP and J48), so the best classifier is clearly BN whatever we use features selection with “Info Gain Attribute Eval” or not, the result is the same.

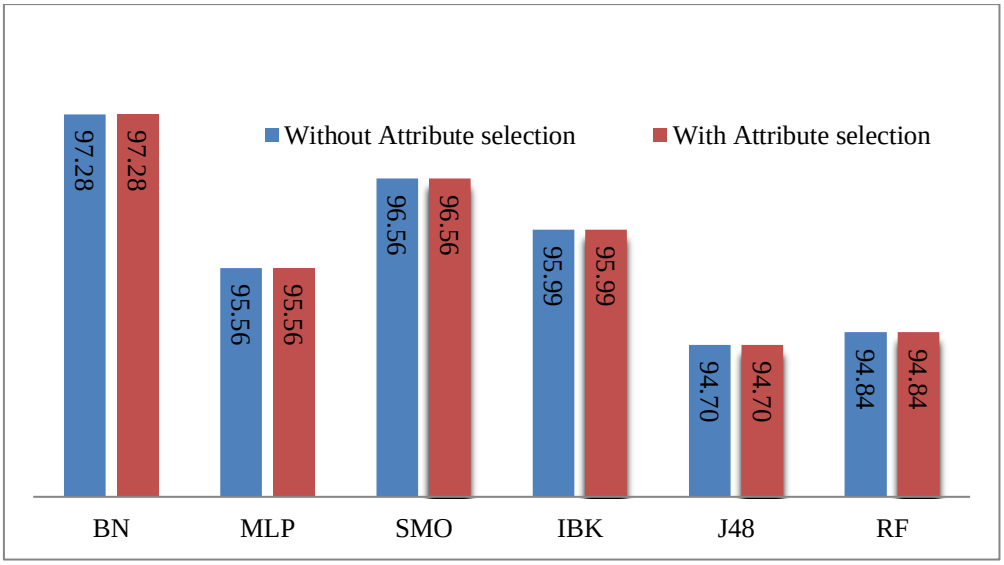

Fig. 4: Single classifier in WBC

ISSN No:-2456-2165

The result of combining BN and each of the other classifiers is shown on Figure 2. As shown as in this figure fusion between BN and RF achieves the best accuracy (98.21%). When using features selection with "Info Gain AttributeEval" on WBC dataset, the accuracy was improved for all combinations of BN with the other classifiers, except "BN-MLP" the accuracy for this combination (67.85%) was decreases.

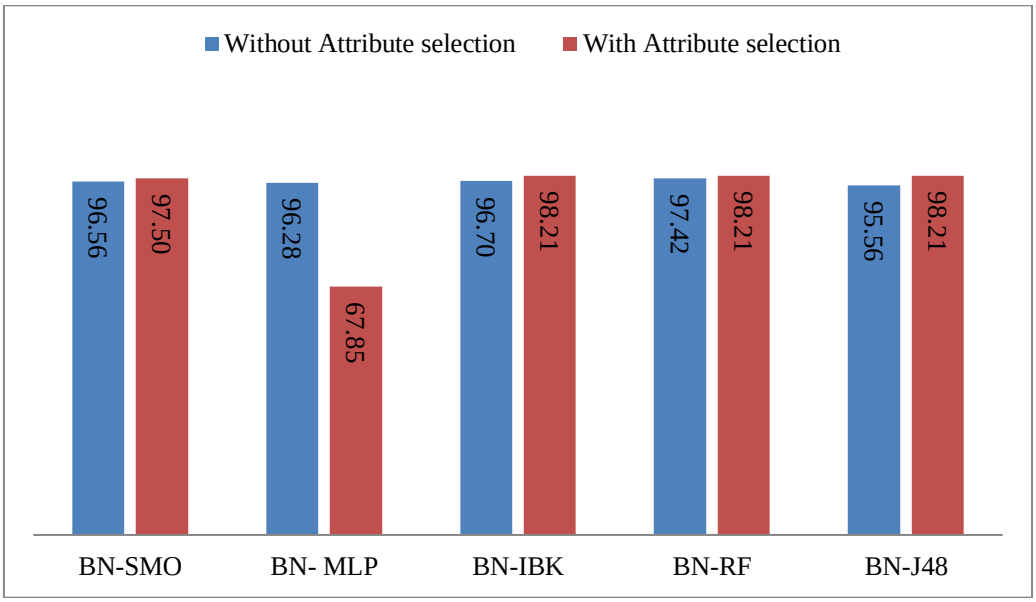

Fig. 5: Fusion of two classifiers in WBC

When we merge three classifiers (classifiers BN+RF+SMO, BN+RF+MLP, BN+RF+J48 and BN+RF+IBK) we conclude that the accuracy decrease to 97.13%; see Figure. 3. In addition in WBC dataset, when we use attribute selection in this case (three classifiers) the accuracy improved for all; however best result (99.64%) is when we merge BN-RF-IBK.

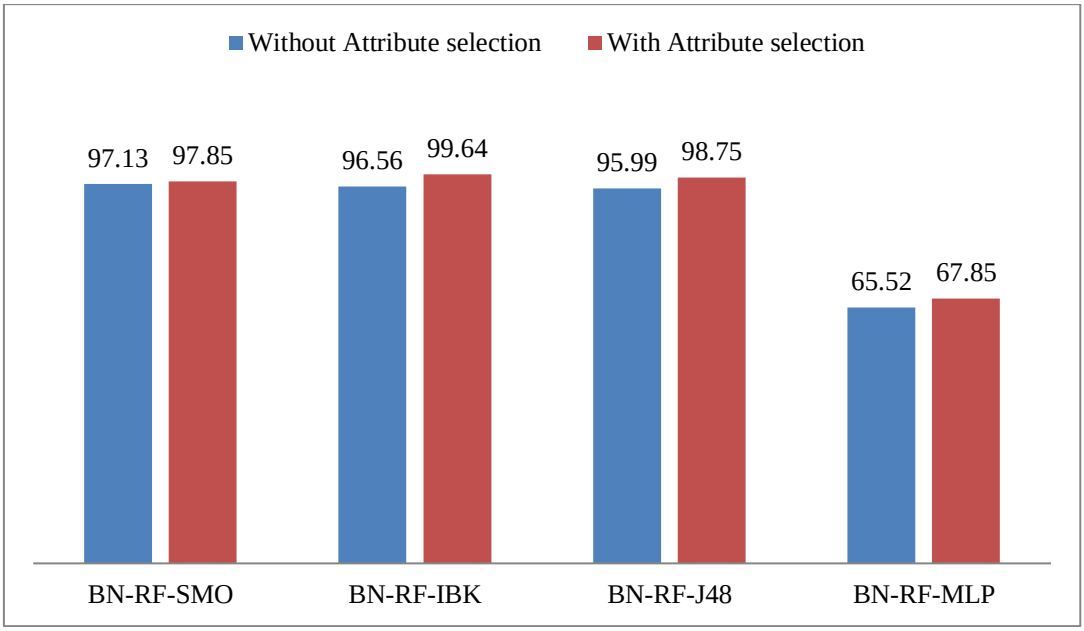

Fig. 6: Combination between three classifiers in WBC

We move on now to merge four classifiers; Figure.4. shows that the fusion between BN, RF, SMO and IBK decrease the accuracy (96.99%). When using features selection with "InfoGainAttributeEval" on WBC dataset, the accuracy was improved for all combinations, the best one was achieved by the combination of BN-RF-IBK-J48 (99.64%).

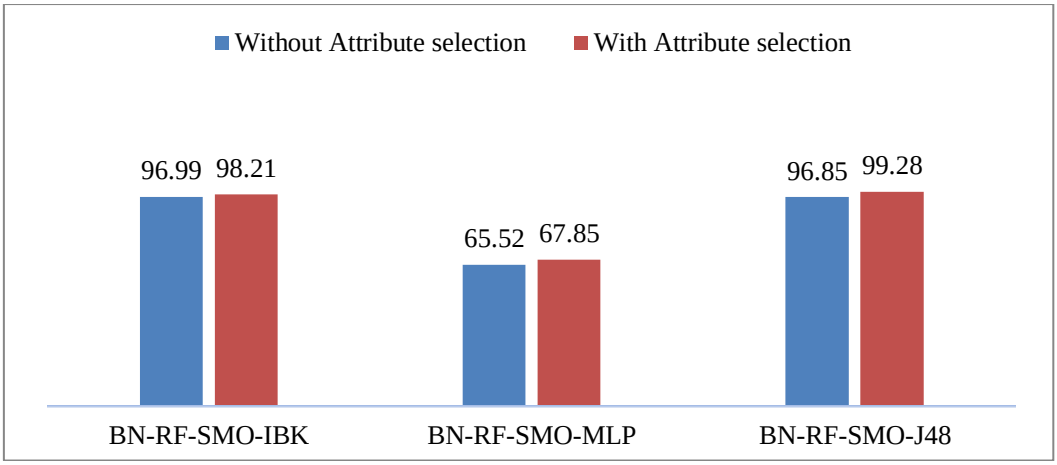

Fig. 7: Fusion of four classifiers in WBC.

These four experiences shown that the fusion of BN-RF-IBK with attribute selection give the best accuracy (99.64%). In the following figure (Fig.5.) we compared classification accuracies of other papers (SVM-RBF Kernel [29], SVM [30], CART [31]) and the recent proposed method for WBC dataset.

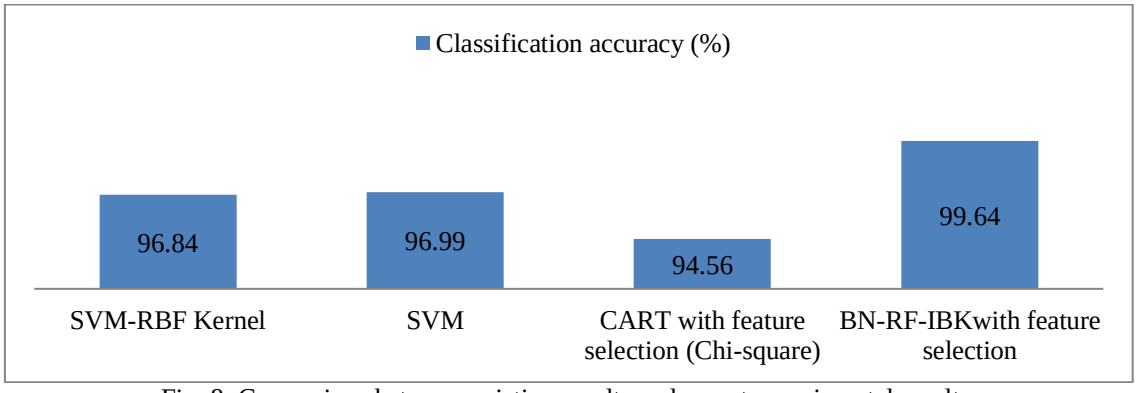

Fig. 8: Comparison between existing results and recent experimental results.

B. Experiment (2) using Wisconsin Diagnosis Breast Cancer (WDBC) dataset with features selection:

In Figure.6. presents the accuracy comparison for the six classifiers (BN, MLP, J48, SMO, RF and IBK) based on supplied test set as a test method. For SMO we have a value of accuracy equal 97.71%, it is more accurate than other classifiers. When we used features selection with "InfoGainAttributeEval" on WDBC dataset, theaccuracy was improved for BN, IBK, J48 and RF and it decreases for MLP and SMO.

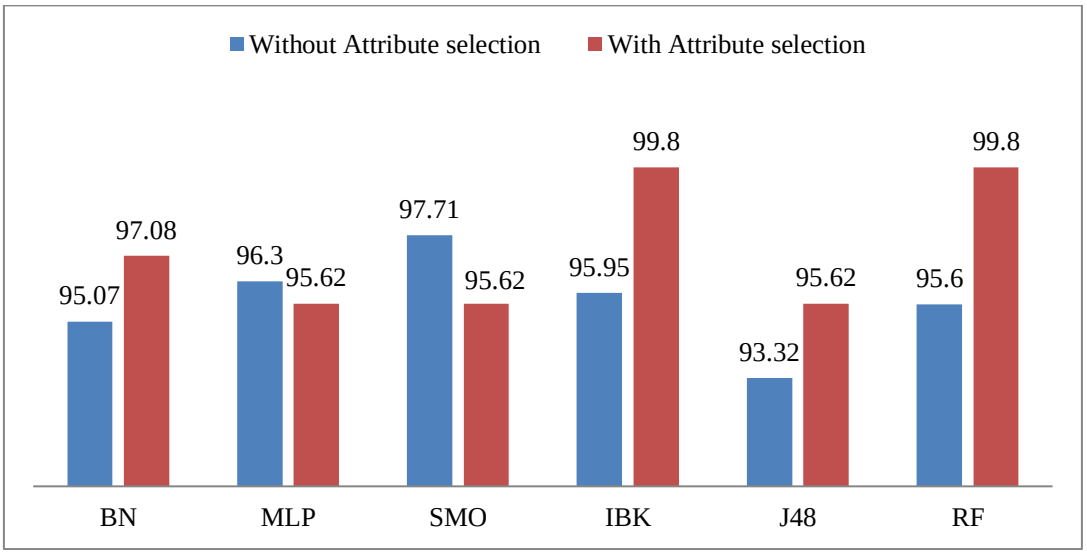

Fig. 9: Single classifier in WDBC

ISSN No:-2456-2165

Figure.7. demonstrates how the fusion of SMO with each of the other classifiers produced the following outcomes: The maximum accuracy (97.81%) is obtained while combining SMO and MLP, SMO and IBK, SMO and BN, and SMO and RF. But if we use features selection with "InfoGainAttributeEval" on WDBC dataset, the accuracy decreases.

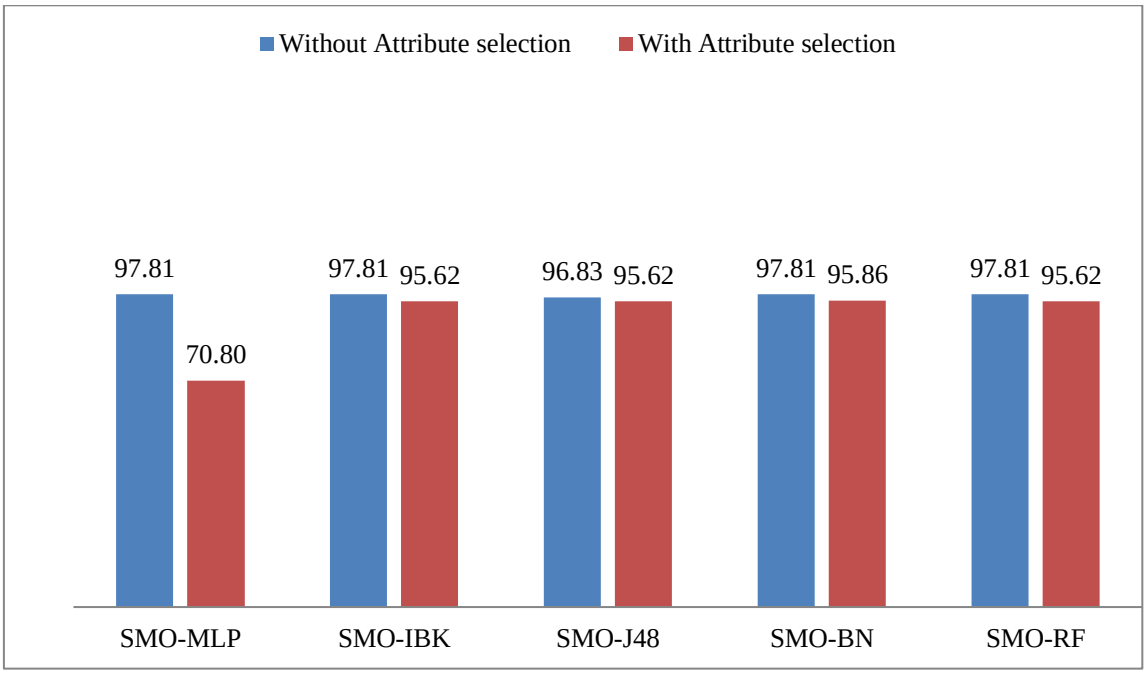

Fig. 10: Fusion of two classifiers in WDBC

Figure .8. Illustrates how the accuracy drops when we attempt to confuse SMO with the other two classifiers. When features selection with "Info Gain Attribute Eval" is used on WDBC dataset, the accuracy increases for all combinations, the best one was achieved by the combination of SMO-RF-IBK (100%).

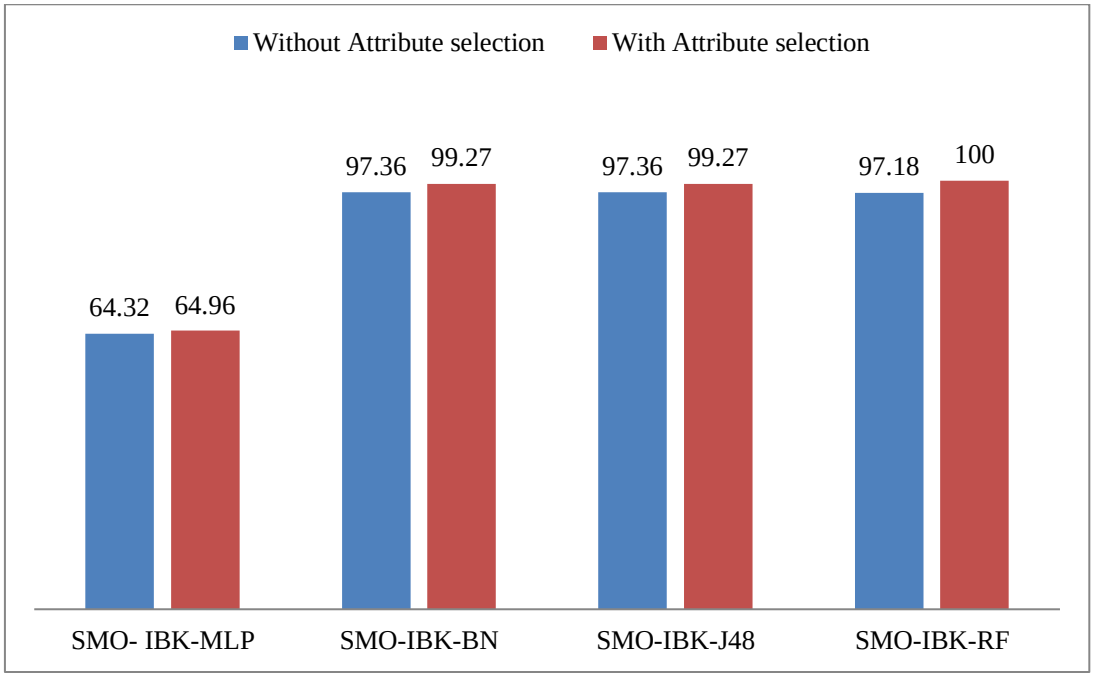

Fig. 11: Fusion of three classifiers in WDBC

Figure.9. shows that the fusion between SMO, IBK and NB with MLP increases the accuracy slightly but still lower than the highest accuracy in single classifiers and fusion of two classifiers. When using features selection with "InfoGainAttributeEval" on WDBC dataset, the accuracy increases for all combinations, the best one was achieved by the combination of SMO- IBK-BN-RF and SMO- IBK-BN-J48 (99.27%).

ISSN No:-2456-2165

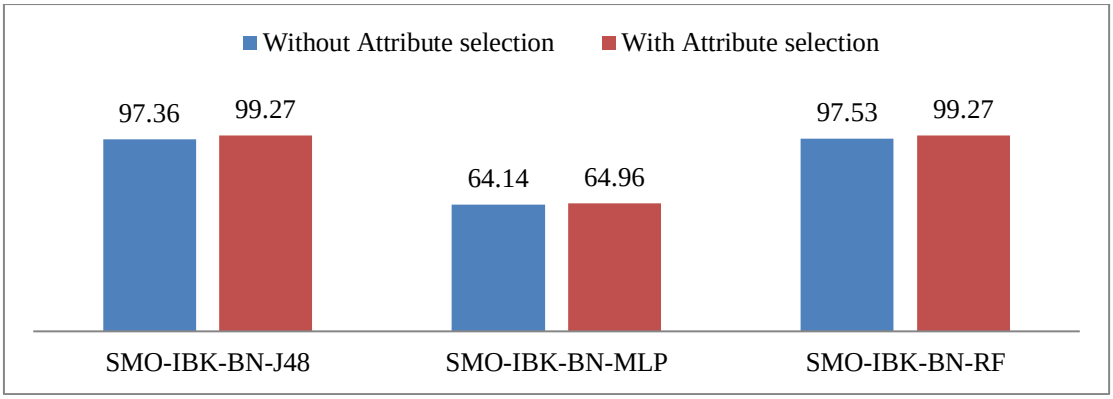

Fig. 12: Fusion of four classifiers in WDBC

Figure.10. provides a comparison of other works' classification accuracy (CART with feature selection (Chi-square) [31], C4.5 [32], Hybrid Approach [33], linear discreet analysis [34], neuron-fuzzy [35], and supervised fuzzy clustering [36]) and recent proposed method (SMO-IBK-RF with attribute selection) for WDBC dataset.

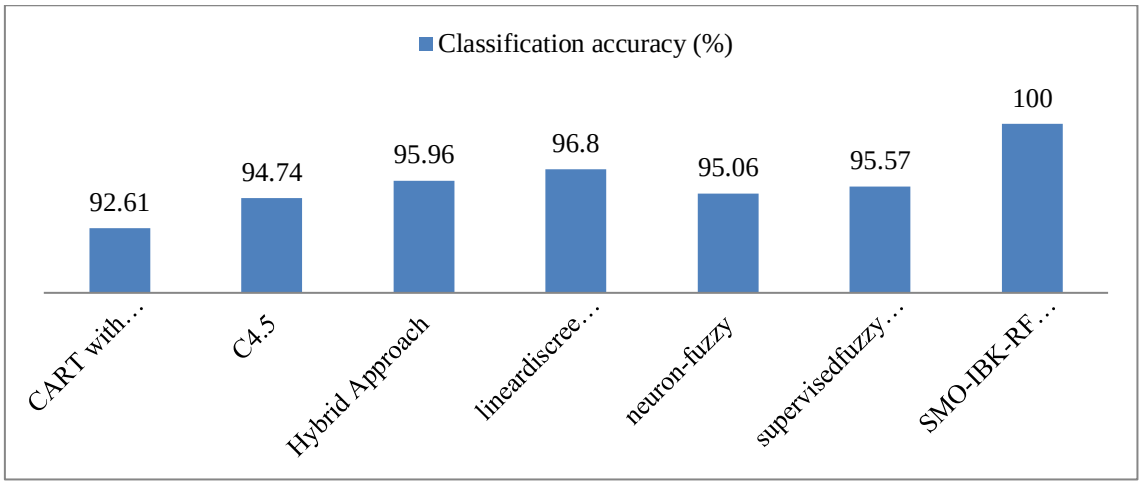

Fig. 13: Comparison of existing and recent experimental results.

C. Experiment (3) using Wisconsin Prognosis Breast Cancer (WPBC) dataset with feature selection

The accuracy comparison of the six classifiers (NB, MLP, J48, SMO, RF, and IBK) using the provided test set as the test method is displayed in Figure. 11. The greatest accuracy is achieved with RF (78.28%). In addition, accuracies of BN and SMO are the same (75.75%). But when we use features selection with "InfoGainAttributeEval" on WPBC dataset, the accuracy was increased for some classifiers and decreased for others. The accuracies of RF (100%) and IBK (100%) are the highest.

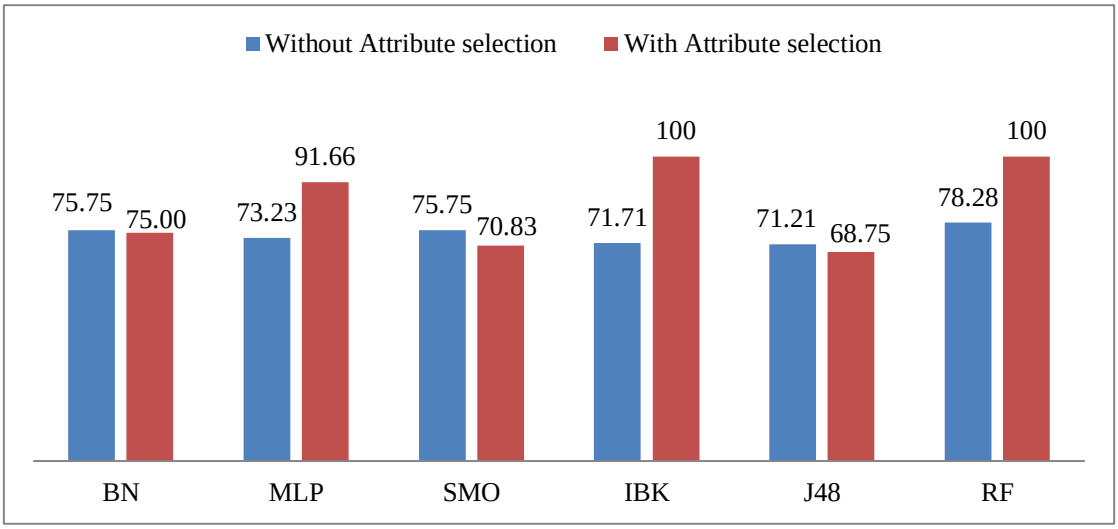

Fig. 14: Single classifier in WPBC

ISSN No:-2456-2165

The fusion of RF with each of the other classifiers produced the following results, as illustrated in Figure12: The best accuracy is obtained when RF and BN fuse together (79.79%), followed by RF and MLP fused together (77.27%). The combination of RF and SMO results in the reduced accuracy. When using features selection with "InfoGainAttributeEval", the accuracy was increased for the fusions (RF-BN, RF-MLP, and RF-J48) and decreased for (RF-SMO). The accuracy of fusion between RF and IBK (100%) is the highest.

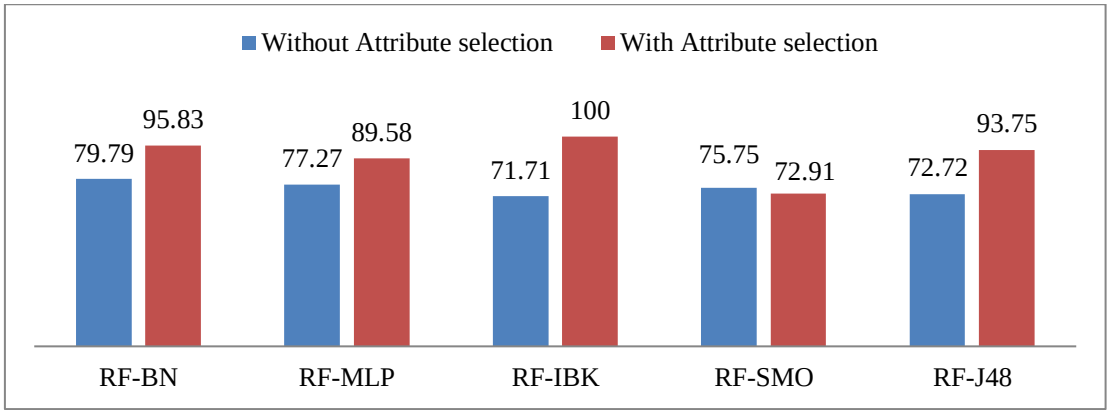

Fig. 15: Fusion of two classifiers in WPBC

The best accuracy of 76.26% is achieved by fusing RF, BN, and MLP, as shown in Figure13. However, this accuracy is less than that of single classification and fusion between two classifiers. The accuracy rises with the use of features selection; the fusion accuracy between RF BN and IBK (98.98%) is the highest.

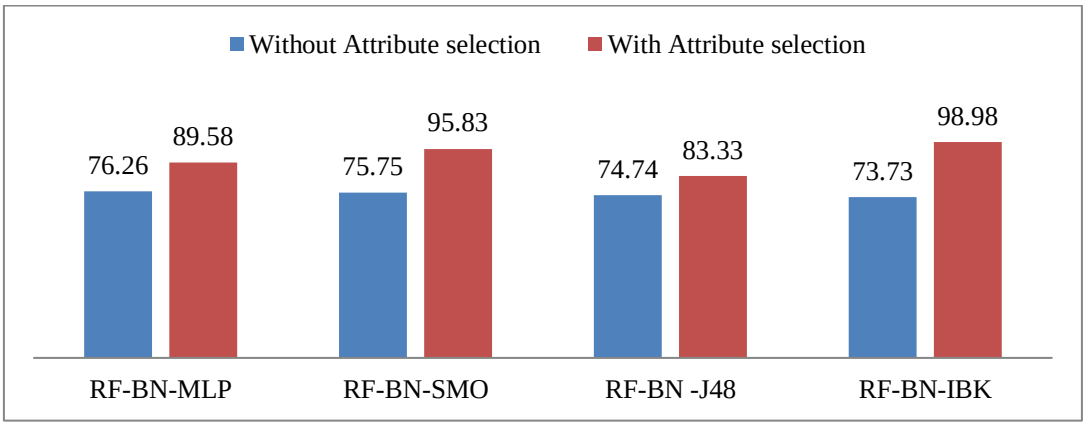

Fig. 16: Fusion of three classifiers in WPBC

Figure 14 shows that the fusion accuracy value between RF, BN, MLP, and SMO is higher than that of other classifiers. Without feature selection, it achieves 77.77%, and with feature selection, it achieves 97.91%.

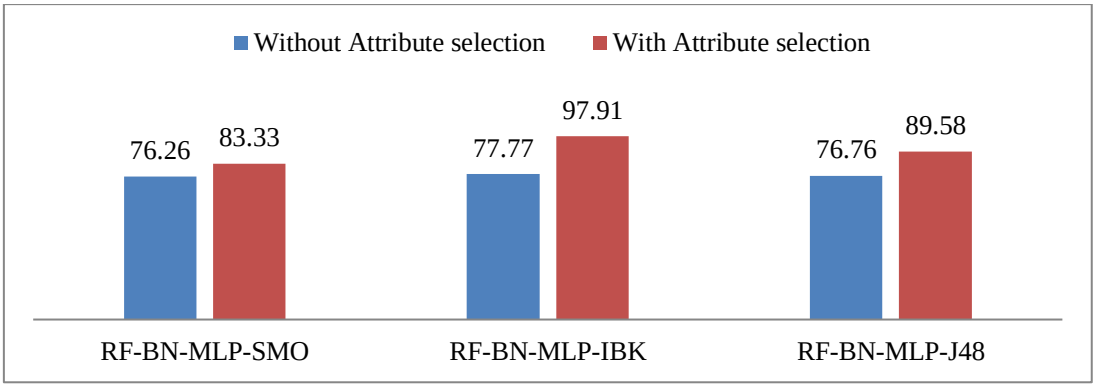

Fig. 17: Fusion of four classifiers in WPBC

A comparison of the recent proposed method for the WPBC dataset with the classification accuracies (ANN [37] and SMO+48+MLP+IBK [38]) of other papers is presented in Figure. 15. It's clear that our proposed method give the best accuracy compared with other methods.

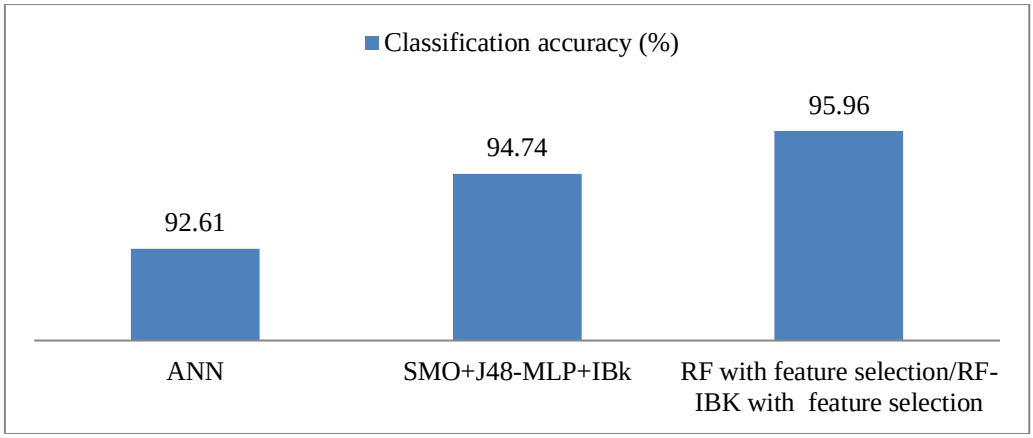

Fig. 18: Comparison of existing and recent experimental results

V. CONCLUSION

In this article we proposed different methods in three different datasets (WBC, WDBC, WPBC), experimentally for WBC dataset the best method of classification is to combine BN, RF and IBK with attribute selection to get the best accuracy, for WDBC dataset the proposed method (fusion of SMO-IBK-RF with attribute selection) give an accuracy of 100 %, In the third experience when we use a WPBC datasets our proposed method surpass the ANN method and the confusion of SMO, J48, MLP and IBK.

REFERENCES

[1]. https://www.contrelecancer.ma/site_media/uploaded_files/Guide_de_detection_pre%C3%BCcoce_des_cancers_du_sein_et_du_col_de_lute%C3%BCrus.pdf .
[2]. Training health workers in clinical breast examination for early detection of breast cancer in low- and middle-income countries, Shahin Sayed, Anthony K Ngugi, Nicole Nwosu, Miriam C Mutebi, Powell Ochieng, Aruyaru S Mwenda, Rehana A Salam, Authors' declarations of interest; 18 April 2023.
[3]. Tsehay Admassu Assegie College of Engineering and Technology, Department of Computing Technology, Aksum University, Aksum, Ethiopia,An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.
[4]. Kaya, S., Yağanoğlu, M. (2020). An Example of Performance Comparison of Supervised Machine Learning Algorithms Beforeand After PCA and LDA Application: Breast Cancer Detection.
[5]. S. B. Sakri, N. B. Abdul Rashid and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," in IEEE Access, vol. 6, pp. 29637-29647, 2018, doi: 10.1109/ACCESS.2018.2843443.
[6]. Halima Bouden, Somya Arach. (2023) ,Breast cancer diagnosis model using multi-classifier fusion. Indian journal of applied research. Volume 13. Issue 10. PRINT ISSN No 2249 - 555X . Yavuz, E., Eyupoglu, C., Sanver, U., & Yazici, R. (2017). An ensemble of neural networks for breast cancer diagnosis. 2017 International Conference on Computer Science and Engineering (UBMK). doi:10.1109/ubmk.2017.8093456.
[7]. Mekha, P., Teeyasuksaet, N. (2019). Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells. Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical,Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), pp. 343-346, doi: 10.1109/ECTINCON.2019.8692297.
[8]. Basunia, M.R., Pervin, I.A., Al Mahmud, M., Saha, S., Arifuzzaman, M. (2020). On Predicting and Analyzing Breast Cancerusing Data Mining Approach. IEEE Region 10 Symposium (TENSYMP), pp. 1257-1260, doi:10.1109/TENSYMP50017.2020.9230871.
[9]. Duda, R.O., Hart, P.E.: “Pattern Classification and Scene Analysis”, In: WileyInterscience Publication, New York (1973).
[10]. Aeinfar, V., Mazdarani, H., Deregeh, F., Hayati, M., & Payandeh, M. (2009). Multilayer Perceptron Neural Network with supervised training method for diagnosis and predicting blood disorder and cancer. 2009 IEEE International Symposium on Industrial Electronics.
[11]. Yavuz, E., Eyupoglu, C., Sanver, U., & Yazici, R. (2017). An ensemble of neural networks for breast cancer diagnosis. 2017 International Conference on Computer Science and Engineering (UBMK). doi:10.1109/ubmk.2017.8093456.
[12]. Bishop, C.M.: “Neural Networks for Pattern Recognition”. Oxford University Press, New York (1999).
[13]. Angeline Christobel. Y, Dr. Sivaprakasam (2011). An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer Information Systems, Vol. 3, No. 2, 2011.
[14]. Tsehay Admassu Assegie College of Engineering and Technology, Department of Computing Technology, Aksum University, Aksum, Ethiopia, an optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.
[15]. Ross Quinlan, (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.
[16]. Dai, B., Chen, R.-C., Zhu, S.-Z., & Zhang, W.-W. Using Random Forest Algorithm for Breast Cancer

Diagnosis. 2018 International Symposium on Computer, Consumer and Control (IS3C).2018
[17]. Vapnik, V.N., The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag, New York, 1995
[18]. K. Srikanth, S. Zahoor Ul Huq, A.P. Siva Kumar, Analysis, Implementation and Comparison of Machine Learning Algorithms on Breast Cancer Dataset using WEKA Tool, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S, March 2019.
[19]. Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSRTR-98- 14, Microsoft Research, 1998.
[20]. Bharati, S., Rahman, M. A., & Podder, P. (2018). Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA. 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT).
[21]. J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kauffman Publishers, 2000.
[22]. NIAKSU, Olegas. CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing, 2015, vol. 3, no 2, p. 92.
[23]. Rdiger Wirth. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, page 2939, 2000.
[24]. HALL, Mark, FRANK, Eibe, HOLMES, Geoffrey, et al. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009.
[25]. A.Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, (2010).
[26]. Špečkauskienė, V. and Lukoševičius, A. (2009) Methodology of adaptation of data mining methods for medical decision support: Case study. Data Mining, 9(2), 228- 235.
[27]. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998). Discovering Data Mining: From Concept to Implementation, Upper Saddle River, N.J., Prentice Hall.
[28]. S. Aruna et al. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer.
[29]. Angeline Christobel. Y, Dr. Sivaprakasam (2011). An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer Information Systems,Vol. 3, No. 2, 2011.
[30]. D.Lavanya, Dr.K.Usha Rani,..," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.
[31]. E.Osuna, R.Freund, and F. Girosi, "Training support vector machines: Application to face detection". Proceedings of computer vision and pattern recognition, Puerto Rico pp. 130–136.1997.
[32]. Lavanya, "Ensemble Decision Tree Classifier for Breast Cancer Data," International Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24, Feb. 2012.
[33]. B.Ster, and A.Dobnikar, "Neural networks in medical diagnosis: Comparison with other methods." Proceedings of the international conference on engineering applications of neural networks pp. 427–430. 1996.
[34]. T.Joachims, Transductive inference for text classification using support vector machines. Proceedings of international conference machine learning. Slovenia. 1999.
[35]. J.Abonyi, and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers." Pattern Recognition Letters, vol.14(24), 2195–2207,2003.
[36]. Vaibhav Narayan Chunekar, Hemant P. Ambulgekar (2009). Approach of Neural Network to Diagnose Breast Cancer on three different Data Set. International Conference on Advances in Recent Technologies in Communication and Computing. 2009.
[37]. Breast cancer diagnosis on three different datasets using multi-classifiers GI Salama, M Abdelhalim, MA Zeid.International Journal of Computer and Information Technology (2277 – 0764) Volume 01–Issue 01, September 2012.