1.1 Introduction :-

Breast cancer is one of the most common cancers with a high mortality rate among women. With the early diagnosis and accurate of breast cancer survival will increase from 56% to more than 86%. Therefore, an accurate and reliable system is necessary for the early diagnosis of this cancer.[1]

Data mining is the survey of large datasets to getting hidden and patterns, relationships and knowledge that are not easy to detect with traditional statistical methods .Knowledge discovery and data mining are concepts that are applied in business more than a decade. By development of data mining technology, it is not only exclusive applied in commercial purposes, but also successfully applied in many different field like medical tasks, for examples in intensive care medicine analysis, time dependency patterns mining in clinical pathways, breast cancer screening and diagnosis of heart disease . Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data.[2]

Data mining used in health can have tremendous potential and usefulness. However, the success of healthcare data mining depends on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. [11] through Mammography .

Mammography is considered the most reliable method in early detection of breast cancer. Due to the high volume of mammograms to be read by physicians, the accuracy rate tends to decrease, and automatic reading of digital mammograms

becomes highly desirable. It has been proven that double reading of mammograms (consecutive reading by two physicians or radiologists) increased the accuracy, but at high costs. That is why the computer aided diagnosis systems are necessary to assist the medical staff to achieve high efficiency and effectiveness.

In data mining breast cancer research has been one of the important field topics in medical science In recent period. The classification of Breast Cancer data can be useful to predict the result of some diseases or discover the genetic behavior of tumors. There are many techniques to predict and classification breast cancer pattern. This work empirically compares performance of different classification rules that are suitable for direct interpretability of their results.

Breast cancer is becoming a cause of death among women in the whole world, meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. In this research work, a decision intelligence technique based ensemble method is proposed for breast cancer diagnosis. In the proposed ensemble, the issue of model selection and feature selection is solved using WEKA

An ensemble include a set of individually trained classifiers (such as neural networks or decision trees) whose predictions are combined when classifying novel instances. Previous research has shown that an ensemble is often more accurate than any of the single classifiers in the ensemble. Bagging and Boosting are two relatively new but popular methods for producing ensembles.

.

1.2 Research background

Automatic diagnostic systems are an important application of analysis of database and pattern recognition, aiming at assisting physicians in marking diagnostic decisions. Automated diagnosis is especially used to diagnose the variety of cancers. Classification is the organization of data in given classes. The most important part in classification approaches or model is the classification algorithm which used to learn from the training set and build the model

Breast cancer classification divides into categories according to different schemes, each based on different criteria and serving a different purpose. The major categories are the histopathological type, the grade of the tumor, the stage of the tumor, and the expression of proteins and genes. As knowledge of cancer cell biology develops these classifications are updated . The purpose of classification is to select the best treatment. The effectiveness of a specific treatment is demonstrated for a specific breast cancer (usually by randomized, controlled trials). That treatment may not be effective in a different breast cancer.

Data are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources. Low-quality data will lead to low-quality mining results. Data preprocessing improve quality of the data and mining results .

1.3 Problem statement

One of the main reasons for the low of accuracy of the diagnosis of breast cancer, is a result of using either not accurate features or not a proper classifier method. Recently most of the researchers state that using more than one classifier is quite better than using a single classifier [8]. This study applied ensemble classification method based on baggin and boosting approach for Breast Cancer detection instead of individual classification. The study also use an important sub features only instead of all data set features based on Correlation Feature Selection (CFS) algorithms.

1.4 Objectives of the Study

1- To solve missing values problem in the (Wisconsin Breast Cancer) data set based common used in the class.
2-To select the most important features only based CFS algorithm.
3- To Apply ensemble method for mammogram image (Wisconsin Breast Cancer) data set.
4-Evaluate the proposed method and compare by the previous studies.

1.5 Significant of the study

Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones.

1.6 Scope of Study

This search covers the offline not online classification and considers The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used to differentiate benign (noncancerous) from malignant (cancerous) samples. The evaluated measures that will used in this thesis is the Confusion Matrix (true

positive, true negative, False Positive and False negative) to determine and examine the accuracy of the classifier that is used during the study.

1.7 Thesis Organization

Chapter one a general definition about data mining and its functionality also describe the problem statement of the study and objective, significant and the scope of the study. Chapter two Literature review of classification method, medical image classification. Chapter three describe the research methodology, the three phases of the thesis and materials that used in the study. Chapter four Describes the implementation of the ensemble classifier, build the classifier and run it in training data, test it after that in test data to determine the accuracy of the classifier. Chapter five gives the conclusion and recommendation of the study.

Chapter (2)

2.1 Introduction

The Data mining is the technique to discover the knowledge which is hidden in the large data sets. It involves with different methods and algorithms to perform efficient analysis over the data sets. The classification is the technique, which is used to mine the data and helps to make the prediction about the future. Different data mining algorithms are available for classification, like C4.5, Simple Cart, Navie Bayesen, Logistic Regression and Multi-Layer Perceptron based on Artificial Neural Network.[3]

2.2 Feature selection methods

Feature selection methods have been used in the machine learning domain. The main aim of these techniques is to remove irrelevant or redundant features from the dataset. Feature selection methods have two categories: wrapper and filter. The wrapper evaluates and selects attributes based on accuracy estimates by the target learning algorithm. Using a certain learning algorithm, wrapper basically searches the feature space by omitting some features and testing the impact of feature omission on the prediction metrics. The feature that make significant difference in learning process implies it does matter and should be considered as a high quality feature. On the other hand, filter uses the general characteristics of data itself and work separately from the learning algorithm. Precisely, filter uses the statistical correlation between a set of features and the target feature. The amount of correlation between features and the target variable determine the importance of target variable [15,18]. Filter based approaches are not dependent on classifiers and usually faster and more scalable than wrapper based methods. In addition, they have low computational complexity.

2.2.1 Information Gain

Information gain (relative entropy, or Kullback-Leibler divergence), in probability theory and information theory, is a measure of the difference between two probability distributions. It evaluates a feature X by measuring the amount of information gained with respect to the class (or group) variable Y, defined as follows:

$$I(X) = H\,(P(Y)-H\,(P(Y/X)) \quad (1)$$

Specifically, it measures the difference the marginal distribution of observable Y assuming that it is independent of feature X(P(Y)) and the conditional distribution of Y assuming that is dependent of X (P(Y/X)). If X is not differentially expressed, Y will be independent of X, thus X will have small information gain value, and vice versa 19].

2.2.2 B. Relief

Relief-F is an instance-based feature selection method which evaluates a feature by how well its value distinguishes samples that are from different groups but are similar to each other. For each feature X, Relief-F selects a random sample and k of its nearest neighbors from the same class and each of different classes. Then X is scored as the sum of weighted differences in different classes and the same class. If X is differentially expressed, it will show greater differences for samples from different classes, thus it will receive higher score (or vice versa) [19].C. One-R. it is a simple algorithm proposed by Holte [20]. It builds one rule for each attribute in the training data and then selects the rule with the smallest error. It treats all numerically valued features as continuous and uses a straightforward method to divide the range of values into several disjoint intervals. It handles missing values by treating “missing” as a legitimate value. This is one of the most primitive schemes. It produces simple rules based on one feature only.

Although it is a minimal form of classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes [16].

2.2.3 . Principal Component Analysis (PCA)

The aim of PCA is to reduce the dimensionality of dataset that contains a large number of correlated attributes by transforming the original attributes space to a new space in which attributes are uncorrelated. The algorithm then ranks the variation between the original dataset and the new one. Transformed attributes with most variations are saved; meanwhile discard the rest of attributes. It's also important to mention that PCA is valid for unsupervised data sets because it doesn't take into account the class label [15, 21].

2.2.4. Consistency Based Subset Evaluation (CS)

CS adopts the class consistency rate as the evaluation measure. The idea is to obtain a set of attributes that divide the original dataset into subsets that contain one class majority [8]. One of well known consistency based feature selection is consistency metric proposed by Liu and Setiono [12].

$$\text{Consistency} = \frac{\sum_{j=0}^{k} |Dj| - |Mj|}{N} \quad (3)$$

where s is feature subset, k is the number of features in s, |Dj| is the number of occurrences of the jth attributes value ,combination, |Mj| is the cardinality of the majority class for the jth attribute's value, and N is the number of features in the original dataset [12].

2.2.5. Correlation Based Feature Selection (CFS)

CFS is a simple filter algorithm that ranks feature subsets and discovers the merit of feature or subset of features according to a correlation based heuristic evaluation function. The purpose of CFS is to find subsets that contain features that are highly correlated with the class and uncorrelated with each other. The rest of

features should be ignored. Redundant features should be excluded as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. CFS feature subset evaluation function is shown as follows [14]:

$$\text{Merits} = \frac{\text{krcf}}{\sqrt{\text{k+(k}-1)\text{rff}}} \qquad (2)$$

where Merits is the heuristic "merit" of a feature subset S containing k features, rcf is the mean feature-class correlation (f € s), and rff is the average feature-feature intercorrelation. This equation is, in fact, Pearson‟s correlation, where all variables have been standardized. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. The heuristic handles irrelevant features as they will be poor predictors of the class. Redundant attributes are discriminated against as they will be highly correlated with one or more of the other features [13].

2.3 General Approach of Classification

The data mining ccontains of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most data mining methods commonly used for this review are of classification category as the applied prediction techniques assign patients to either a ”benign” group that is non- cancerous or a ”malignant” group that is cancerous and generate rules for the same. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems.

In data mining, classification is one of the most important task. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups[4].

2.3.1. Decision Trees (DT's)

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item.

Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached.

Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

2.3.2. Support Vector Machine(SVM)

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

2.3.3. Genetic Algorithms (GAs) / Evolutionary Programming (EP)

Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution of collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one

expects that the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism.

2.3.4 Fuzzy Sets

Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

2.3.5. Neural Networks

Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.[9].

2.3.6 Ensemble Classifiers

Ensemble Classifiers: An ensemble classifier combines several classifiers is known as multiple classifier. In general, an ensemble classifier performance is higher than single classifier performance. The ensembles are used to improve the classification performance of a single classifier. However, not all ensemble classifiers were performing better than single classifier. Three popular ensemble methods are bagging, multiboost and random subspace has been applied to various

machine learning algorithms. Ensemble methods have been used to improve the prediction accuracy by combining an ensemble of weak classifiers.

Several methods of estimation have preceded boosting approach. Common feature for all methods is that they work out by extracting samples of a set, calculating the estimate for each drawn sample group repeatedly and combining the calculated results into unique one. One of the ways, the simplest one, to manage estimation is to examine the statistics of selected available samples from the set and combine the results of calculation together by averaging them. Such approach is a jack-knife estimation, when one sample is left out from the whole set each time to make an estimation . Obtained collection of estimates is averaged afterwards to give the final result. Another, improved method, is Bootstrapping. Bootstrapping repeatedly draws certain number of samples from the set and processes calculated estimations by averaging, similar to jack-knife .

2.3.6.1 Bagging

Bagging is the further step towards boosting. It consists of Bootstrap aggregation which increases classifier stability and reduces variance over a collection of samples. In this, samples are drawn with replacement and each draw has a classifier Ci attached to it, so that final classifier becomes a weighted vote of Ci - s. Bootstrapping and Bagging techniques are non-adaptive Boosting techniques[8**].**

2.3.6.1 Boosting

Boosting is a general method for improving the performance of any learning algorithm. The method works by repeatedly running a weak learner, on various distributed training data. The classifiers produced by the weak learners are then

combined into a single composite strong classifier in order to achieve a higher accuracy than the weak learner's classifiers would have had.

The main idea of this algorithm is to assign a weight in each example in the training set. In the beginning , all weights are equal, but in every round, the weights of all misclassified instances are increased while the weights of correctly classified instances are decreased. As a on sequence, the weak learner is forced to focus on the difficult instances of the training set. This procedure provides a series of classifiers that complement one another.

2.4 Medical image classification

Medical Imaging is becoming an essential component in various fields of bio-medical research and clinical practice: Neuroscientists detect regional metabolic brain activity from positron emission tomography (PET), functional magnetic resonance imaging (MRI), and magnetic resonance spectrum imaging (MRSI) scans, biologists study cells and generate 3D confocal microscopy data sets, virologists generate 3D reconstructions of viruses from micrographs, and radiologists identify and quantify tumors from MRI and computed tomography (CT) scans.

Image classification techniques help to detect subjects suffered from particular diseases and to detect disease-related regions. Medical image Classification can play an important role in diagnostic and teaching purposes in medicine. For these purposes different imaging modalities are used. There are many classifications created for medical images using both grey-scale and color medical images. One way is to find the texture of the images and have the analysis [23].

2.5 RELATED WORKS :-

number	paper name	Date	Technique	Results	Open issues
1	Critical Analysis of Data Mining Techniques on Medical Data	2016	talked about the latest technologies data mining in medical Data	found an appropriate decision tree with a cancer diagnosis has obtained a 93%	to improve the accuracy that we use bagging or boosting technology to increase accuracy.
2	Data Mining Techniques: To Predict and resolve Breast Cancer Survivability	2015	RepTree, RBF Network and Simple Logistic.	Simple logistic Classification with accuracy of 74.47%	can use other features such as breast images, genetic marker features, smoking
3	A Survey on Breast Cancer Analysis Using Data Mining Techniques	2014	Fuzzy association rules and neural network,	The medical data processed Missing ,incomplete	Medical data suffer other problems not addressed in the paper, such as inconsistent
4	Breast Cancer Risk Prediction Using Data Mining Classification Techniques	2014	naïve bayes, J48	J48 94.2 , naïve bayes 82.6	Just used 69 instance
5	Diagnosis of Brest cancer using decision tree data mining technique	2015	J48 decision tree	94 %	Can using bagging or boosting to improve the accuracy

Chapter (3)

3.1 introductions

The classification process in this study involves three major steps namely Data preprocessing is it include feature selection, classification and evaluation. Here, we use ensemble based classifier method, more details about research phase is explained below:

3.2 Research phases

This study emphasis of three phases starting in Data preprocessing, classification and end with testing and evaluating.

3.2.1 Phase1 Data Preprocessing

Databases are highly vulnerable to noisy, missing and inconsistent data due to their typically massive size and their likely origin from multiple, miscellaneous sources. Hence data preprocessing is a necessary phase for classification purposes. Data preprocessing includes data cleaning, data dimensionality reduction, data transformation (data normalization, data binning) followed by classification.

Here, the data cleaning technique includes removing the missing values if present, with the most commonly used of the attributes. Data normalization brings the range of all attribute values between 0 and 1, and using CFS algorithms as a feature selection technique.

Dataset used in this study downloaded from the UCI Machine Learning Repository [10] . Has been converted to the Data Set formation excel to csv format compatible with Weka tool and the sample number is not required in the formation of the model it is removed from the record and J48 requires its class label to be nominal (String) in type. Missing values are processed by the most commonly used value for Bare Nuclei There are 16 cases Missing Value and

Discretization data from 1 – 10 to 1-3 (reduces and simplifies the original data), also in this phase CFS feature selection algorithm was applied and the most important features selection accordingly .

3.2.2 Phase 2 Classification :-

Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. In the second stage, the classification model constructed previously is used to classify unknown classes' data which is known as a testing.

3.2.2.1 Ensemble classifier

This classifier consist of select a collection (ensemble) of hypotheses and combine their predictions. For example generate 100 different decision trees from the same or different training set and have them vote on the best classification for a new example. It has methods usage to classify

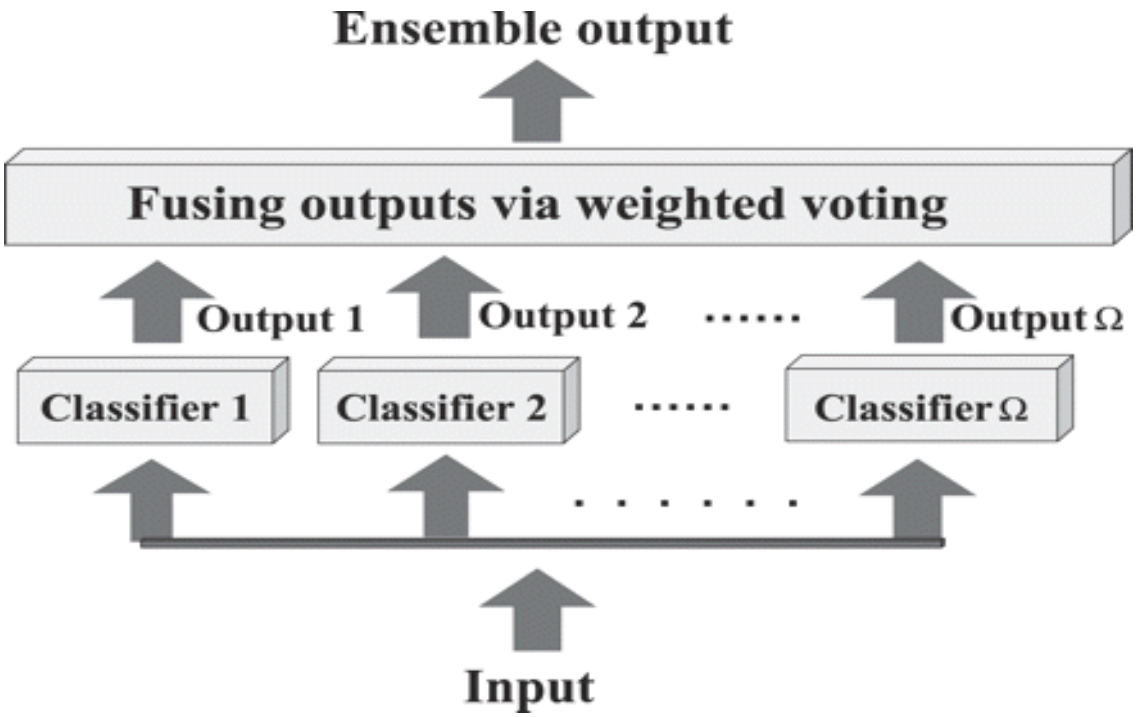

Fig No 3.1: Ensemble classifier

3.2.2.2Building classifier

The idea of building a predictive model by integrating multiple models has been under investigation for a long time. the history of ensemble methods starts as early as 1977 with Tukeys Twicing, an ensemble of two linear regression models.

3.2.2.3Method of ensemble classifier

Ensemble methods can be used for improving the quality and robustness of clustering algorithms. The most common types of ensemble methods are:

3.2.2.3.1 Bagging

Bagging works as a method of increasing accuracy. to demonstration that , Suppose that you are a patient and would like to have a diagnosis made based on your symptoms. Instead of asking one doctor, you may choose to ask several. If a certain diagnosis occurs more than any of the others, you may choose this as the

final or best diagnosis. That is, the final diagnosis is made based on a majority vote, where each doctor gets an equal vote. Now replace each doctor by a classifier, and you have the basic idea behind bagging. Intuitively, a majority vote made by a large group of doctors may be more reliable than a majority vote made by a small group. Given a set, *D*, of *d* tuples, bagging works as follows. For iteration *i* (*i* = 1, 2, : : : , *k*), a training set, *Di*, of *d* tuples is sampled with replacement from the original set of tuples, *D*. Note that the term bagging stands for bootstrap aggregation. Each training set is a bootstrap sample. Because sampling with replacement is used, some Algorithm: Bagging. The bagging algorithm create an ensemble of models (classifiers or predictors) for a learning scheme where each model gives an equally-weighted prediction. The algorithm is summarized below

Input:

- *D*, a set of *d* training tuples;
- *k*, the number of models in the ensemble;
- a learning scheme (e.g., decision tree algorithm, backpropagation, etc.)

Output: A composite model, *M**.

Method:

(1) for *i* = 1 to *k* do // create *k* models:

(2) create bootstrap sample, *Di*, by sampling *D* with replacement;

(3) use *Di* to derive a model, *Mi*;

(4) end for

To use the composite model on a tuple , ***X***:

(1) if classification then

(2) let each of the *k* models classify ***X*** and return the majority vote;

(3) if prediction then

(4) let each of the *k* models predict a value for ***X*** and return the average predicted value;

Of the original tuples of *D* may not be included in *Di*, whereas others may occur more than once. A classifier model, *Mi*, is learned for each training set, *Di*. To classify an unknown tuple, ***X***, each classifier, *Mi*, returns its class prediction, which counts as one vote. The bagged classifier, *M**, counts the votes and assigns the class with the most votes to ***X***. Bagging can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple. The bagged classifier often has significantly greater accuracy than a single classifier derived from *D*, the original training data. It will not be considerably worse and is more robust to the effects of noisy data. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers. For prediction, it was theoretically proven that a Bagging predictor will always have improved accuracy over a single predictor derived from *D*.

3.2.2.3.2Boosting

Return to the previous example ,suppose that as a patient, you have certain symptoms. Instead of consulting one doctor, you choose to consult several. Suppose you assign weights to the value or worth of each doctor's diagnosis, based on the accuracies of previous diagnoses they have made. The final diagnosis is then a combination of the weighted diagnoses. This is the essence behind boosting. In boosting, weights are assigned to each training tuple. A series of *k* classifiers is iteratively learned. After a classifier *Mi* is learned, the weights are updated to allow the subsequent classifier,*Mi*+1, to "pay more attention" to the training tuples that were misclassified by *Mi*. The final boosted classifier, *M**, combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy. The boosting algorithm can be extended for the prediction of continuous values. Ada boost is a popular boosting algorithm. Suppose we would

like to boost the accuracy of some learning method. We are given *D*, a data set of *d* class-labeled tuples, (***X*1**, *y*1), (***X*2**, *y*2), … (***Xd***, *yd*),where *yi* is the class label of tuple ***Xi***. Initially , Adaboost assigns each training tuple an equal weight of 1/*d*. Generating *k* classifiers for the ensemble requires *k* rounds through the rest of the algorithm. In round *i*, the tuples from *D* are sampled to form a training set, *Di*, of size *d*. Sampling with replacement is used—the same tuple may be selected more than once. Each tuple's chance of being selected is based on its weight. A classifier model, *Mi*, is derived from the training tuples of *Di*. Its error is then calculated using *Di* as a test set. The weights of the training tuples are then adjusted according to how they were classified(Previous classified). If a tuple was incorrectly classified, its weight is increased. If a tuple was correctly classified, its weight is decreased. A tuple's weight reflects how hard it is to Classify the higher the weight, the more often it has been misclassified. These weights will be used to generate the training samples for the classifier of the next round. The basic idea is that when we build a classifier, we want it to focus more on the misclassified tuples of the previous round. Some classifiers may be better at classifying some "hard" tuples than others. In this way, we build a series of classifiers that complement each other. Now, let's look at some of the math that's involved in the algorithm. To compute the error rate of model *Mi*, we sum the weights of each of the tuples in *Di* that *Mi* misclassified. That is,

$$error(Mi) = \sum_{j=0}^{d} \text{Wj} * err(\boldsymbol{Xj}) \quad (3.2).$$

where *err*(***Xj***) is the misclassification error of tuple ***Xj***: If the tuple was misclassified, then *err*(***Xj***) is 1. Otherwise, it is 0. If the performance of classifier *Mi* is so poor that its error exceeds 0.5, then we abandon it. Instead, we try again by generating a new *Di* training set, from which we derive a new *Mi*. The error rate of *Mi* affects how the weights of the training tuples are updated. If a tuple in round *i* was correctly classified, its weight is multiplied by *error*(*Mi*)=

(1-*error*(*Mi*)).Once the weights of all of the correctly classified tuples are updated, the weights for all tuples (including the misclassified ones) are normalized so that their sum remains the same as it was before. To normalize a weight, we multiply it by the sum of the old weights, divided by the sum of the new weights. As a result, the weights of misclassified tuples are increased and the weights of correctly classified tuples are decreased, as described above.

"Once boosting is complete, how is the ensemble of classifiers used to predict the class label of a tuple, X?" Unlike bagging, where each classifier was assigned an equal vote

Algorithm: Adaboost. A boosting algorithm create an ensemble of classifiers. Each one gives a weighted vote.

The algorithm is summarized below .

Input:

- *D*, a set of *d* class-labeled training tuples;
- *k*, the number of rounds (one classifier is generated per round);
- a classification learning scheme.

Output: A composite model.

Method:

(1) initialize the weight of each tuple in *D* to 1=*d*;

(2) for *i* = 1 to *k* do // for each round:

(3) sample *D* with replacement according to the tuple weights to obtain *Di*;

(4) use training set *Di* to derive a model, *Mi*;

(5) compute *error*(*Mi*), the error rate of *Mi* (Equation 3.2)

(6) if *error*(*Mi*) > 0.5 then

(7) reinitialize the weights to 1=*d*

(8) go back to step 3 and try again;

(9) endif

(10) for each tuple in *Di* that was correctly classified do

(11) multiply the weight of the tuple by *error*(*Mi*)=(1-*error*(*Mi*)); // update weights

(12) normalize the weight of each tuple;

(13) endfor

To use the composite model to classify tuple, ***X***:

(1) initialize weight of each class to 0;

(2) for *i* = 1 to *k* do // for each classifier:

(3) *wi* = *log* 1-*error*(*Mi*)/*error*(*Mi*) ; // weight of the classifier's vote

(4) *c* = *Mi*(***X***); // get class prediction for ***X*** from *Mi*

(5) add *wi* to weight for class *c*

(6) endfor

(7) return the class with the largest weight;

Boosting assigns a weight to each classifier's vote, based on how well the classifier performed. The lower a classifier's error rate, the more accurate it, and therefore, the higher its weight for voting should be. The weight of classifier Mi's vote is log (1-error(Mi)/error(Mi)) (3.2) For each class, c, we sum the weights of each classifier that assigned class c to X. The class with the highest sum is the "winner" and is returned as the class prediction for tuple X."How does boosting compare with bagging?" Because of the way boosting focuses on the misclassified tuples, it risks overfitting the resulting composite model to such data.Therefore, sometimes the resulting "boosted" model may be less accurate than a single model derived from the same data. Bagging is less susceptible to model overfitting. While both can significantly improve accuracy in comparison to a single model, boosting tends to achieve greater accuracy.

3.2.3 Phase 3 Testing and Evaluating :-

As far as the classification performance of the model is concerned, the classification rate (C) denotes the percentage of correctly classified samples, which is computed by the following formula.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.3)$$

Where **TP** True positives refer to the positive tuples that were correctly labeled by the classifier, **TN** true negatives are the negative tuples that were correctly labeled by the classifier, **FP** false positives are the negative tuples that were incorrectly, **FN** false negatives are the positive tuples that were incorrectly labeled by the classifier .

3.3 Correlation based Feature Selection (CFS) Algorithm

At the core of the CFS algorithm is a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypothesis on which the heuristic is based can be stated. Good feature subsets contain features highly correlated with (predictive of) the class. yet uncorrelated with (not predictive of) each other.

Feature Correlations Classification tasks in machine learning often involve learning from categorical features as well those that are continuous or ordinal. In order to have a common basis for computing the correlations necessary for equation continuous features are transformed to categorical features in a preprocessing step using the supervised discretisation method .

The purpose of feature selection is to decide which of the initial features to include in the final subset and which to ignore. If there are n possible features

initially then there are n possible subsets. The only way to find the best subset would be to try them all this is clearly prohibitive for all but a small number of initial features.

CFS starts from the empty set of features and uses a forward best first search with a stopping criterion of five consecutive fully expanded non improving subsets.

Figure 3.2 shows the stages of the CFS algorithm and how it is used in conjunction with a machine learning algorithm. A copy of the training data is first discretized ,then passed to CFS. CFS calculates feature-class and feature-feature correlations using symmetrical uncertainty and then searches the feature subset space. The subset with the highest merit (as measured by Equation 2 in chapter 2) found during the search is used to reduce the dimensionality of both the original training data and the testing data. Both reduced datasets may then be passed to a machine learning algorithm for training and testing .

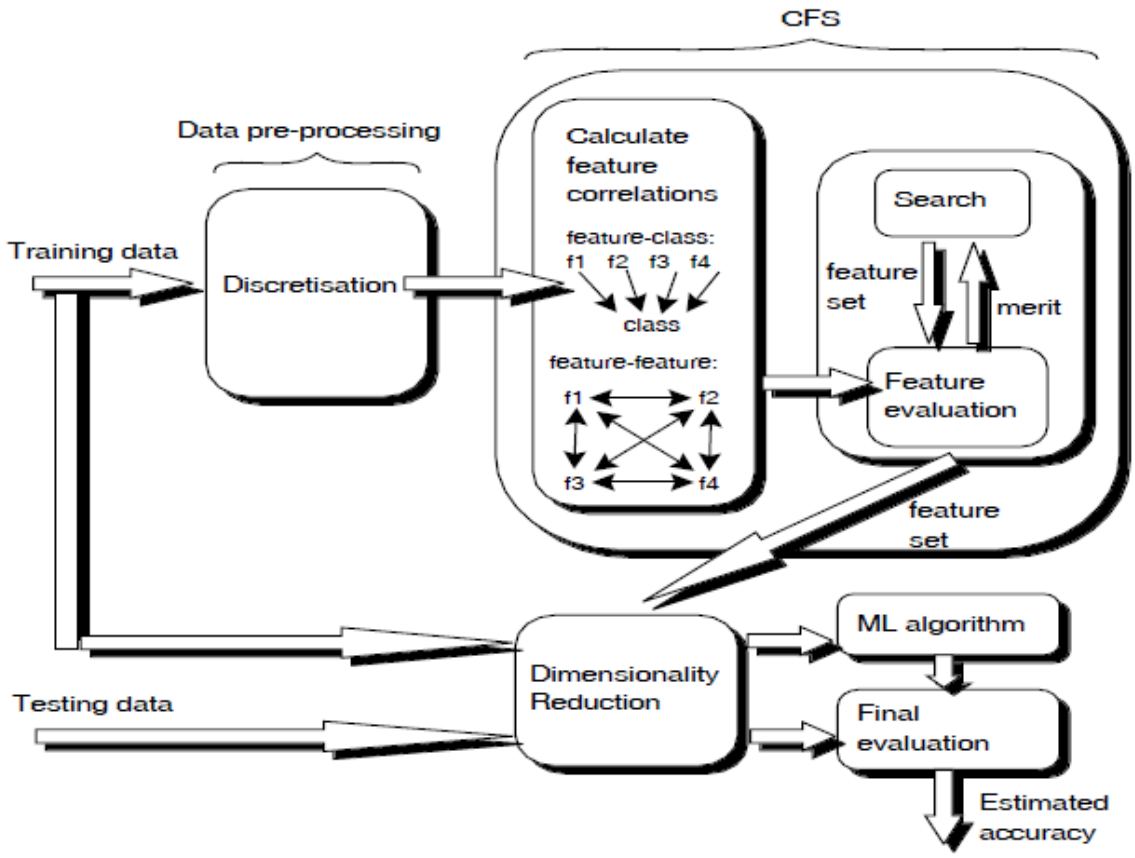

Fig No 3.2 The components of CFS.

3.4 Material and Tools

3.4.1 WEKA (Waikato Environment for Knowledge Analysis)

WEKA is tools for data preparation and research developed at the University of Waikato in New Zealand. When searching for the model that best approximates the target function, it is necessary to provide measures of quality models and learning

3.4.2 Dataset

The Wisconsin Breast Cancer datasets from the UCI MachineLearning Repository [10] is used to differentiate benign (noncancerous) from malignant (cancerous) samples. Table 3.1 shows a brief description of the dataset that is being considered.

Table 3.1. Description of Breast Cancer Dataset

Dataset	**No. Of Attributes**	**No. Of Instances**	**No. fClasses**
Wisconsin Breast Cancer (Original)	**11**	**699**	**2**

Details of the attributes present in the dataset are shown in Table 3.2

Table 3.2 Wisconsin Breast Cancer Dataset Attribute

No	**Attribute**	**Domain**
1	Sample Code Number	Id number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2(Benign) or 4(Malignant)

- **Clump Thickness:** Monolayer grouping in benign and multi layer grouping for cancerous cells.
- **Marginal Adhesion:** Normal cells stick together while cancer cells lose their ability. This is also relating factor to a single epithelial cell size, which is enlarged for a malignant cell.
- **Bare Nuclei:** Benign tumors have nuclei, which are not surrounded by cytoplasm.
- **Bland Chromatin:** Cancer cells have coarse chromatin.
- **Mitoses:** Uncontrollable levels of mitoses (celldivision) are seen in cancer cells.

The dataset comprises of 699 instances of breast cancer patients with each, either having malignant or benign type of tumor. Figure 3.2 shows the distribution of the patient based on the class label (malignant or benign).

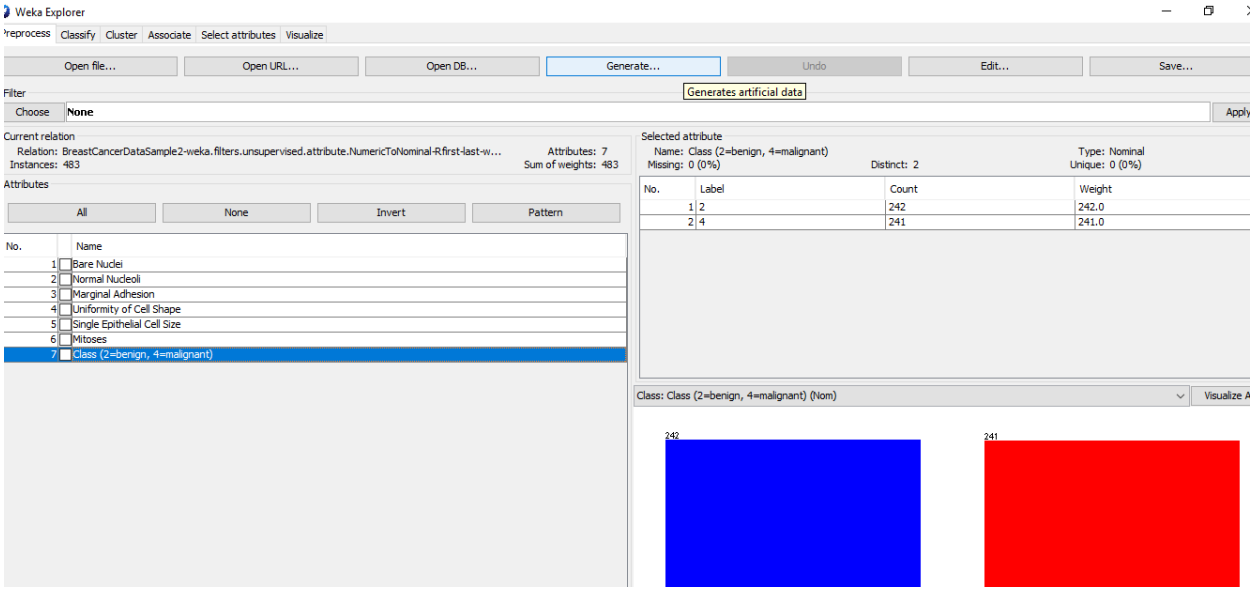

Fig No 3.3 distribution of the patient based on the class label (malignant or benign).

Chapter (4)

4.1 Introduction

In this chapter we use four types of experiments. The first experiment based on individual classifier and all features are taken. The second experiment based on a ensemble method classifier and all features ,The third and fourth experiments are same as previous two experiments , but we use the selected features only based on Correlation Based Feature Selection (CFS) algorithm and apply the pre-processing and also all missing value are solved accordingly.

4.2 Experiment and results

We have used the Weka toolkit in our experiments with the Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version3.7 was utilized as a data mining tool to evaluate the performance and effectiveness of the breast cancer preliminary prediction models and 6 ensemble models were built from several techniques. This is because the WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models. The performance of a chosen classifier is validated based accuracy. The accuracy (AC) is the proportion of the total number of predictions that were correct.

In our first experiment all features were taken, without data processing .and used three data mining algorithms, J84, rep and random forest, as will show below:

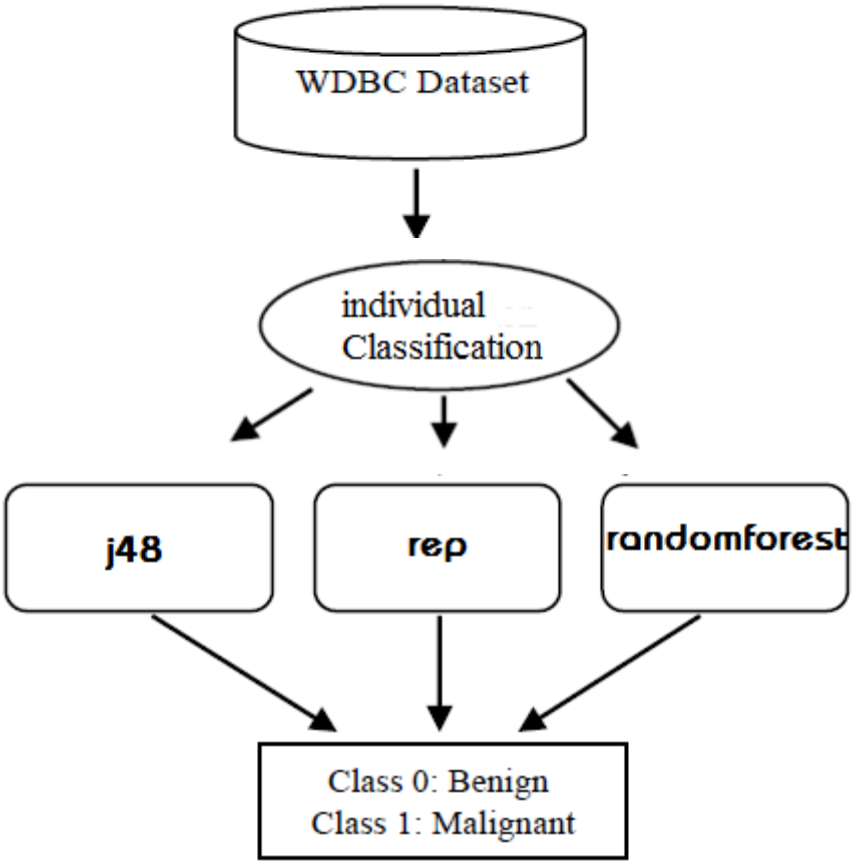

Fig No 4.1: individual classification without data processing .

In the second experiment we apply concept of ensemble classification based on ada boosting and bagging using the three data mining algorithms. J84, rep and random forest as classifier for boosting and, bagging .as shown below:

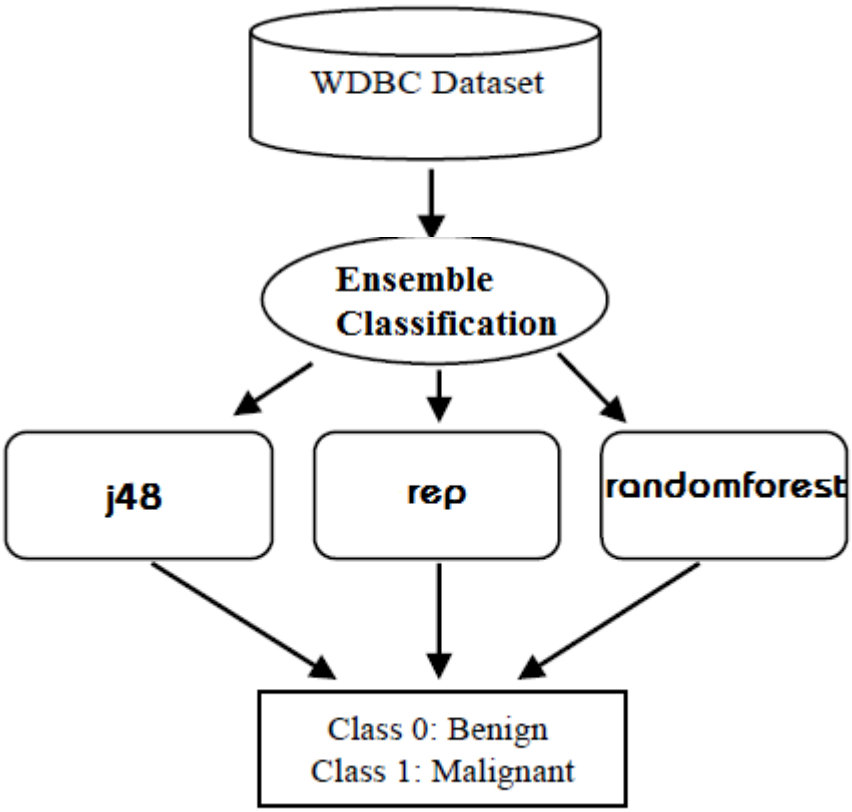

Fig No 4.2 : Ensemble classification using j84,rep and random forest algorithms.

4.3. Results

First without pre-processing for individual classifier and ensembles methods .Also splitting data set to 65% , 70%, 80% as training data set .

Table 4.1 individual and ensemble classification accuracies without pre processing

percent Splitting data	Individual			Baggin Ensemble			Boosting Ensemble		
	J84	**Rep**	**Random**	**J84**	**Rep**	**Random**	**J84**	**Rep**	**Random**
65- 35	90.28	52.57	**95.428**	92	55.428	**96**	92	52.57	**96.571**
70-30	91.33	54	**96**	93.33	52.667	**96.667**	94.66	58	96
80-20	93	57	**96**	95	49	95	93	60	**95**

The following figure 4.3 also give more details of the accuracy for all value.

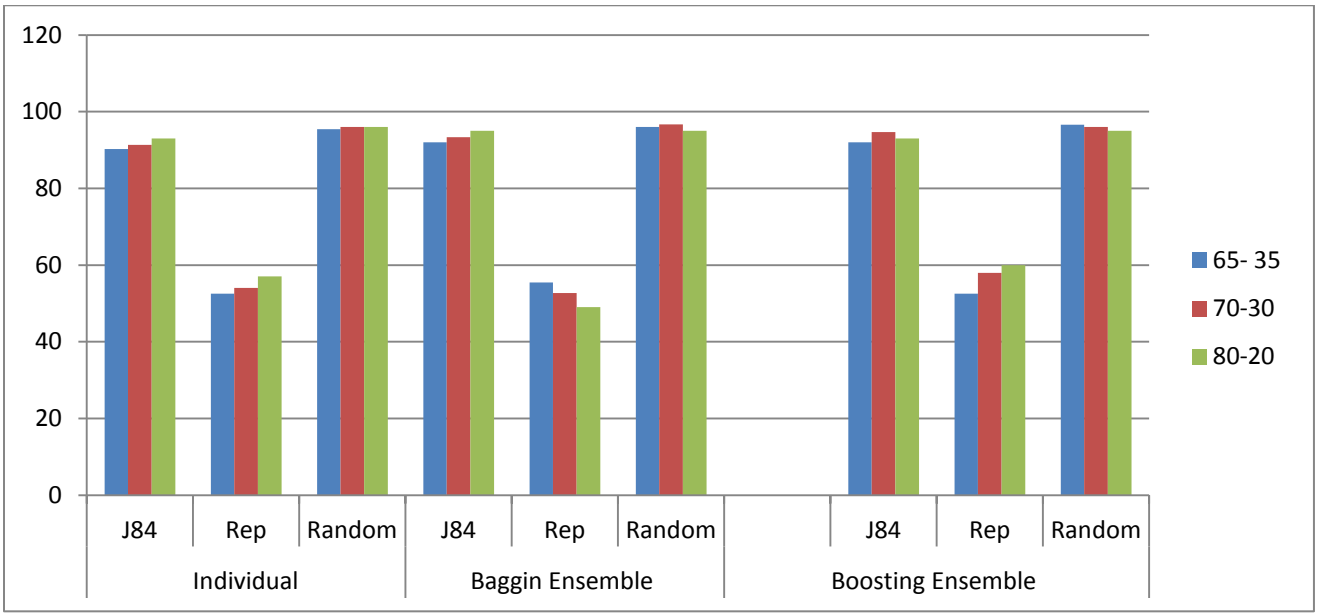

Fig No 4.3: individual and ensemble classification accuracies without pre processing**.**

From the results shown in the previous table and figure, we can see a clear increase in the accuracy of the model when using ensembles methods (ada boosting , bagging) rather than using an individual classifier .

In the third and fourth experiments solving the problem missing value and doing feature selection based on CFS

The most important feature selected by CFS algorithm are shown below

TABEL4.2: features obtain by features selection (CFS)

No	Features
1	Bare Nuclei
2	Normal Nucleoli
3	Marginal Adhesion
4	Uniformity of Cell Shape
5	Single Epithelial Cell Size
6	Mitoses
7	Bland Chromatin
8	Class label

The following figure shows the work flow of the third experiment.

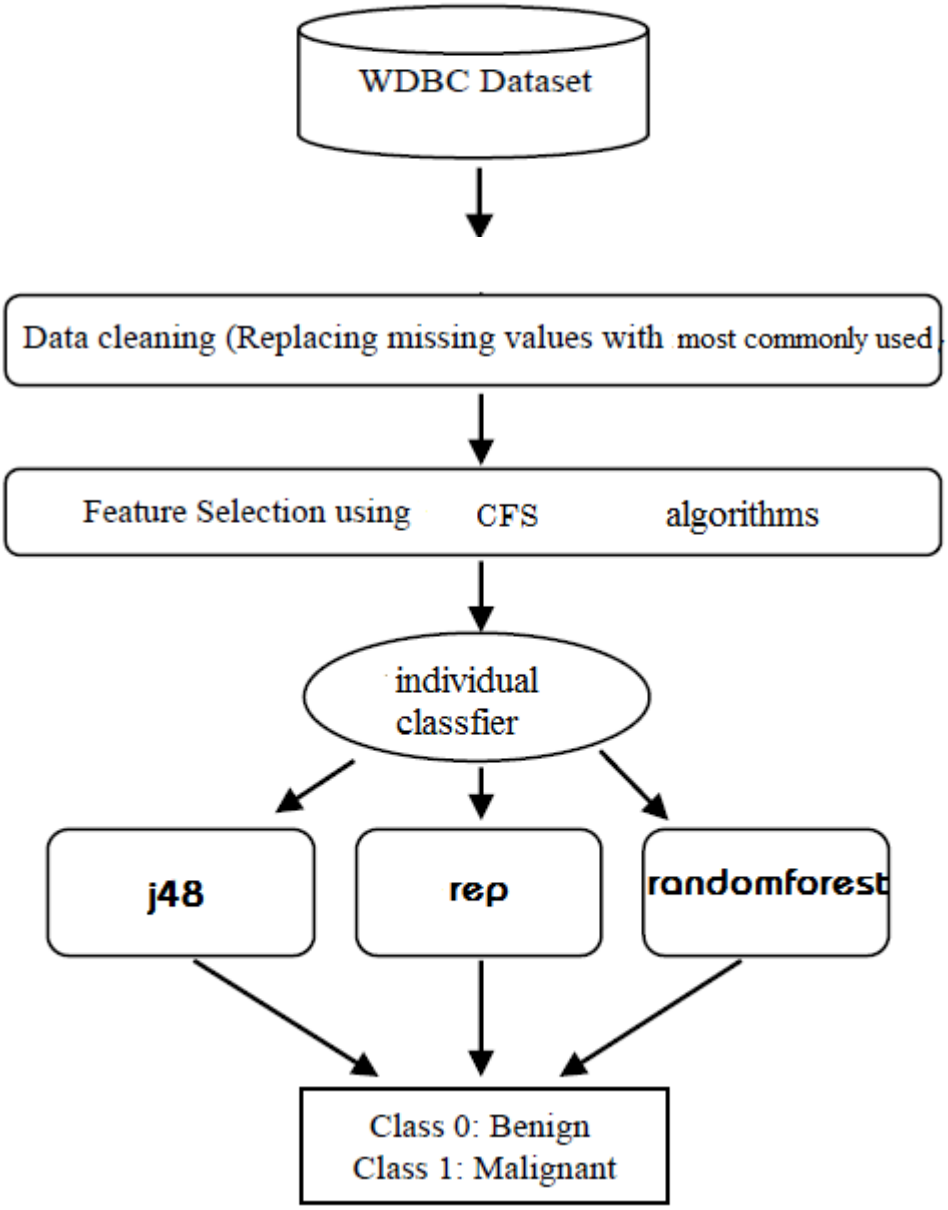

Fig No 4.4: individual classification with pre processing

In the fourth experiment we implemented ada boosting and bagging using the three data mining algorithms. J84, rep and random forest as classifier for boosting, bagging. The following figure shown the work flow of the fourth experiment

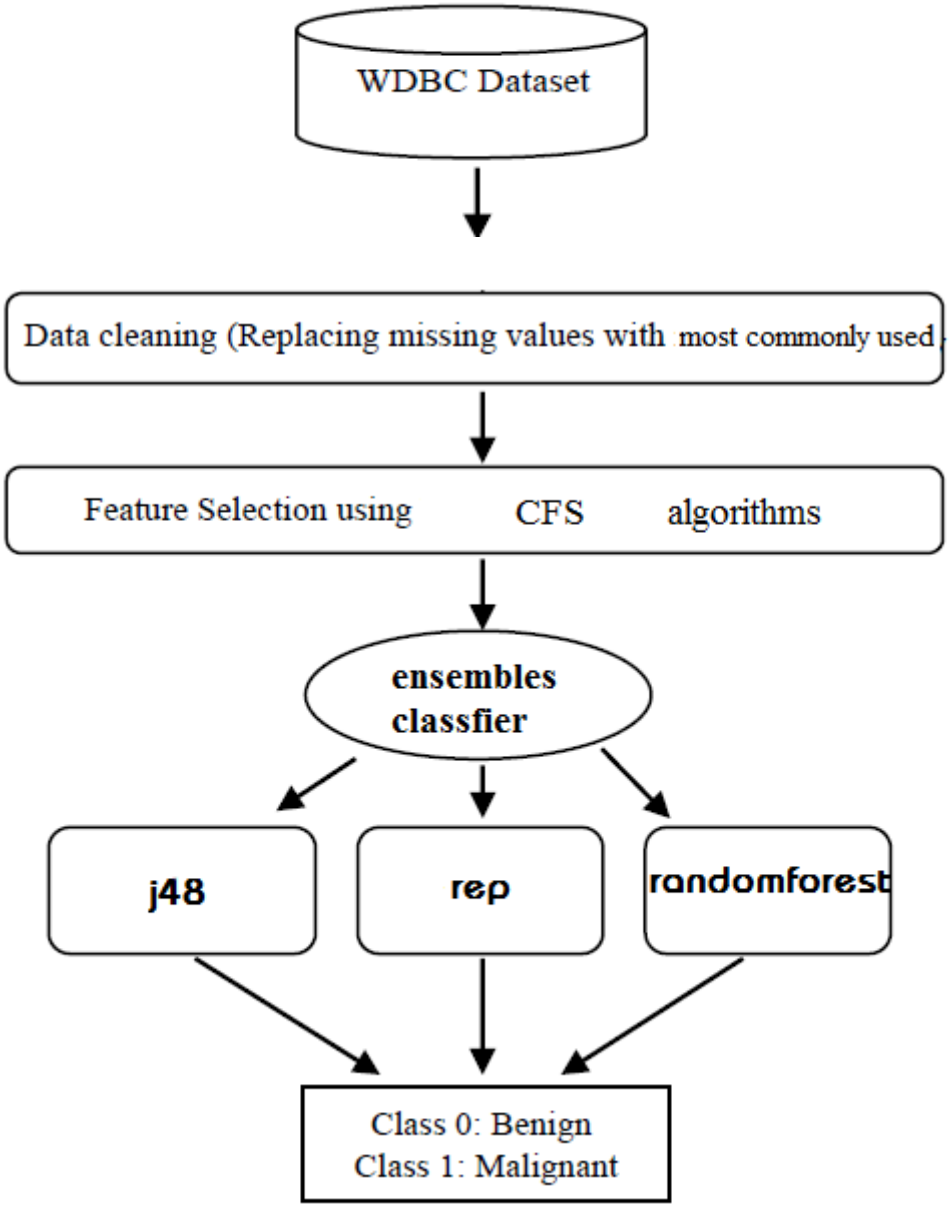

Fig No 4.5: ensemble classification with pre processing

The results of last two with experiment is show below .

Table 4.3 Accuracies of ensemble methods and individual with training data set 65%, 70%, 80%.

percent Splitting data	Individual			Baggin Ensemble			Boosting Ensemble		
	J84	**Rep**	**Random**	**J84**	**Rep**	**Random**	**J84**	**Rep**	**Random**
65- 35	90.2857	90.85	**94.8571**	93.14	93.14	**94.857**	94.86	93.71	**94.285**
70-30	90.667	94	**97.3333**	95.34	94.67	**97.333**	94.67	96	**96.666**
80-20	87	92	**96**	91	93	**96**	96	95	**96**

The following figure 4.6 also gives more details of the accuracy for all values.

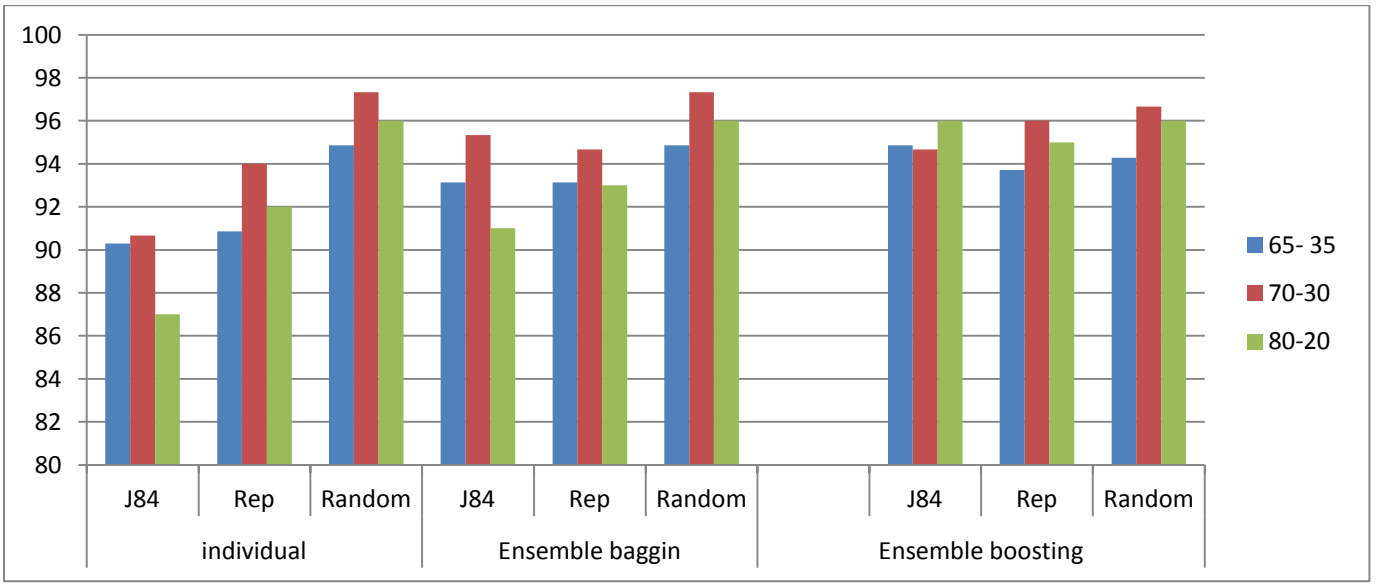

Fig No 4.6 the accuracies of ensemble methods and individual with training data set 65%, 70%, 80%.

Form previous results not that 70% as data set training data set give as better Accuracy than 65 % or 80 % as training data set.

Table 4.3 the accuracies of ensemble methods and individual with pre-processing and 70 % of data set using as training data set .

Algorithm	**J48**	**Rep**	**Random**
individual classifier	90.3667	94	97.3333
Baagin	95.3333	94.6667	97.3333
Boosting	94.67	96	96.3333

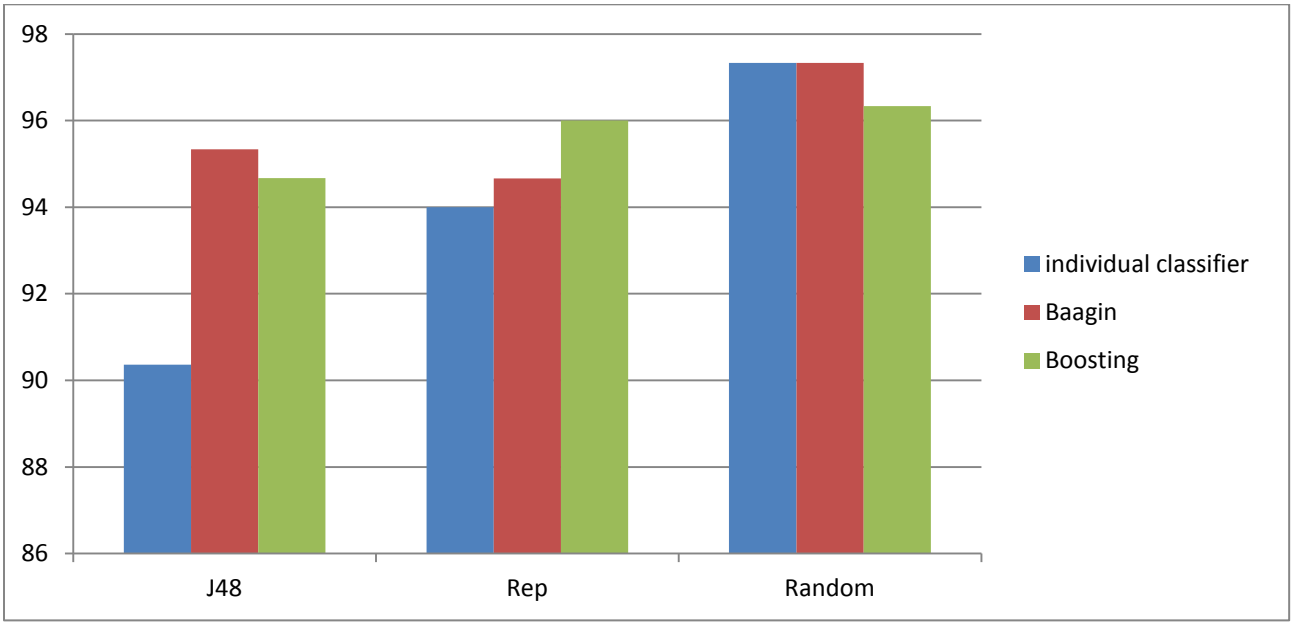

Fig No 4.7: Accuracy of classification methods after apply pre-processing and features selection.

4.4 Discussions

After the end of the four experiments we conclude the following. The first and second experiments, we proved that the Ensemble classification gives better results than the individual classification, although data processing is not used. This can be measured by the arithmetic average of each classification accuracy ,the data division by 70-30 , The individual ,baggin ,boosting by cconsecutively 80.44333333, 80.888, 82.88666667.

In the third and fourth experiment, data pree processing and used feature selection algorithm (CFS) were applied to the first and second experiments, which led to an increase in the accuracy of the classification. This can be measured by the arithmetic mean to individual ,baggin ,boosting by ccconsecutively is 94.0001 , 95.781, 95.77867 .

We note from the previous results improvement in accuracy when using ensembles method and apply pre Processing

4.5 Comparison with other study

In the paper [24] breast cancer prediction was done using Decisions Tree Support Vector Machine Hybrid (DT-SVM) Model. This study was performed using the Wisconsin Breast Cancer Dataset (WBCD) taken as input from UCI machine learning repository (UCI Repository of Machine Learning Databases). The dataset contained 699 instances taken from patients' breasts, of which 458 cases belonged to benign class and the remaining 241 cases belonged to malignant class. It should be noted that there were 16 instances which had missing values. In this study all the missing values were replaced by the mean of the attributes. Each record in the database had nine attributes. These nine attributes were found to

differ significantly between benign and malignant samples. In case of DT-SVM the accuracy obtained was 91% with an error rate of 2.58%. Other classification algorithms had also been applied like IBL (instance-based learning), SMO (Sequential Minimal Optimization) and Naïve Bayes. For IBL the accuracy obtained was 85.23% with an error rate of 12.63%. For SMO the accuracy was 72.56% with an error rate of 5.96%. For Naïve Bayes the accuracy obtained was 89.48% with an error rate of 9.89%. So this comparative study revealed that DT-SVM performed well in classifying the breast cancer data compared to all other algorithms.

In the paper [25] the core objective was to develop a probabilistic breast cancer prediction system using Naive Bayes Classifiers which can be used in making expert decision with highest accuracy. The system may be implemented in remote areas like countryside or rural regions, to imitate like human diagnostic expertise for treatment of cancer disease. The system is user friendly and reliable as model was already developed. For training Wisconsin Datasets containing 699 records with 9 medical attributes was used. For Testing 200 records were taken. This dataset had almost 65.5% benign cases and remaining 34.5% malignant cases. The accuracy was found to be 93%.

In the paper [26] the data set consisted of 699 patient's records of which 499 were considered for training and 200 for testing purposes. Among them, 241 or 34.5% were reported to have breast cancers while the remaining 458 or 65.5% were non-cancerous. In order to validate the prediction results of the six popular data mining techniques the 10-fold crossover validation was used. The k-fold crossover validation was usually used to reduce the error coming from random sampling to compare the accuracies of a number of prediction models. The entire set of data was randomly divided into k folds with the same number of instances in

each fold. The training and testing were performed for k times and one fold was selected for further testing while the rest were selected for further training. The present knowledge distributes the data into 10 folds where 1 fold was used for testing and 9 folds were used for training purpose in the 10-fold crossover validation. Here by applying Naïve Bayes algorithm on testing data an accuracy of 94.5% had been obtained. Same result had been obtained for SVM.

Chapter 5

5.1 CONCLUSIONS

The early detection of breast cancer is needed in reducing life losses. This early breast cancer cell detection can be predicted with the help of modern machine learning techniques . In this study, feature selection (CFS) , ensemble classification techniques have been applied for predicting breast cancer as accurately as possible . This study reveals that ensemble classifier gives the maximum accuracy compared individual classification classifiers

5.2 Recommendation

To enhance the result of the classifier we recommend to use other algorithm as classifier of ensemble method and use other algorithm to selection features .

References

[1] Montazeri, Mitra ,Beigzadeh, Amin (2015)"**Machine learning models in breast cancer survival prediction "** · Technology and health care: official journal of the European Society for Engineering and Medicine

[2] Boshra Bahrami, Mirsaeid Hosseini Shirvani,(2015)**'' Prediction and Diagnosis of Heart Disease by Data Mining Techniques''** Journal of Multidisciplinary Engineering Science and Technology (JMEST)

[3] Renuka Devi , Maria Shyla(2016) " **Critical Analysis of Data Mining Techniques on Medical Data**" International Journal of Applied Engineering Research ISSN 0973-4562 pp 727-730 © Research India Publications. http://www.ripublication.com .

[4] Vikas Chaurasia1, Saurabh Pal2(2014) "**Data Mining Techniques:To Predict and Resolve Breast Cancer Survivability**" , IJCSMC.

[5] Velmurugan T(2014)" **A Survey on Breast Cancer Analysis Using Data Mining Techniques**" researchgate.

[6] Abdulhamit Subasi (2015) "**Breast Cancer Risk Prediction Using Data Mining Classification Techniques**" researchgate.

[7] Ronak Sumbaly(2014) " **Diagnosis of Brest cancer using decision tree data mining technique**". Researchgate

[8] Sateesh Kumar(2012)" **Boosting Techniques on Rarity Mining**" , International Journal of Advanced Research in Computer Science and Software Engineering.

[9] SHELLY GUPTA , DHARMINDER KUMAR , ANAND SHARMA **(2011) "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS"** Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE)

[10] William H Wolberg, Olvi Mangasarian**,** UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].Irvine, CA

[11] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth(1996) "**From data mining to knowledge discovery in databases**". AI magazine,

[12] H. Liu and R. Setiono,(1995) "**CHI2: Feature selection and discretization of numeric attributes**," in *Proc. the 7th IEEE International Conference on Tools with Artificial Intelligence*.

[13] M. A. Hall and L. A. Smith,(1999) "**Feature selection for machine learning:comparing a correlation-based filter approach to the wrapper**,"
Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, pp. 235-239.

[14] M. A. Hall, (1999)"**Correlation-based feature selection for machine learning**," PhD, Department of Computer Science, The University of Waikato, Hamilton

[15] M. Ashraf, G. Chetty, and D. Tran,(2013) "**Feature selection techniques on thyroid,hepatitis, and breast cancer datasets**," *International Journal on Data Mining and Intelligent Information Technology Applications(IJMIA)*,vol. 3, no. 1, pp. 1-8.

[16] J. Novakovic, P. Strbac, and D. Bulatovic(2011), "**Toward optimal feature selection using ranking methods and classification algorithms**," *Yugoslav Journal of Operations Research*, vol. 21, no. 1, pp. 119-135.

[17] H. Yasin, T. A. Jilani, and M. Danish, (2011)"**Hepatitis-C classification using data mining techniques**," *International Journal of Computer Applications*, vol. 24, no. 3, pp. 1-6.

[18] M. Leach, (2012)"**Parallelising feature selection algorithms**," University of Manchester, Manchester.

[19] I. Lee, G. H. Lushington, and M. Visvanathan(2011), "**A filter-based feature selection approach for identifying potential biomarkers for lung cancer**," *Journal of clinical Bioinformatics*, vol. 1, no. 11, pp. 1-8.

[20] R. C. Holte (1993), "**Very simple classification rules perform well on most commonly used datasets,"** *Machine Learning*, vol. 11, pp. 63-91.

[21] I. T. Jolliffe. (2002). Principal Component Analysis. [Online]. Available: http://books.google.com.au

[22] Müller, H., et al. (2012)**"Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks"**. in *CLEF (online working notes/labs/workshop)*.

[23]Animesh Hazra , Subrata Kumar Mandal , Amit Gupta(2016)'' **Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms'' ,** *International Journal of Computer Applications*

[24]K. Sivakami, (2015)"**Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model**." *International Journal of Scientific Engineering and Applied Science (IJSEAS).*

[25] Shweta Kharya, Shika Agrawal and Sunita Soni, *(2014)* "**Naïve Bayes Classifiers: Probabilistic Detection Model for Breast Cancer.**"*International Journal of Computer Applications 92.10.*

[26] G. Ravi Kumar, Dr G. A. Ramachandra and K. Nagamani. *(2013)* "**An Efficient Prediction of Breast Cancer Data using Data Mining Techniques.**" *International Journal of Innovations in Engineering and Technology (IJIET) 2.4).*