Plan Works :

Dedication i
ACKNOWLEDGMENTS ii
ABSTARCT iii
RESUME iv
PLAN WORKS v
LIST OF FIGURES vii
LISTE OF TABLE viii
Glossary ix

INTRODUCTION 1

1 PROBLEMATIC .. 3
1. Background 3
1.1 Intoduction 3
1.2 impact of breast cancer 3
1.3 definition of breast cancer 4
1.4 symptoms of breast cancer 5
1.5 risk factors of breast cancer 5
1.6 stages of breast cancer 6
1.7 treatement 7
2. Diagnostic support 8
3. Problematic 9
4. Conclusion 10

2. Machine learning for breast cancer 11
1. Intoduction 11
2. Supervised classification 11
2.1 Artificial neural networks 12
2.2 Decision trees 14
2.3 Discriminant analysis 14
2.4 k-nearest neighbor 15
2.5 Support vector machines 15
3. Unsupervised classification 15
3.1 Hierarchical clustering 16
3.2 Partitioning clustering 17
4. Selection of feature 18
2.1 Wrapper methods 19
2.2 Filter methods 19

2.3 Embedded methods 20
5. Selection of features and classification in literature 22
6.Recent challenges in breastcancer management 25
7. Conclusion 25

3. Result and Discussion 27
1. Intoduction 27
2. Dataset 27
3.Step of selection 28
3.1 Feature selection using SVM-PSO 28
4. Steup of classification 32
5.Result 33
5.1 Discusion of result 35
6. Summary of the selection techniques 37
7.Conclusion 38

Conclusion 39
Appendix
References

Listes of figures:

1.1 Distribution of mortality of the most common cancers in ALGERIA according to the World Health Organization in 2014.

1.2 The incidence of the most common cancers according to the central register World Health Organization in 2014.

1.3 Normal breast tissue [2].

2.1 Artificial Neural Network [7].

2.2 Decision tree that might be used in breast cancer diagnosis and treatment [8].

2.3 Support Vector Machines [19].

2.4 Hierarchical clustering [23].

2.5 Partitioning clustering [24].

2.6 The wrapper model [32].

2.7 The Filter model [32].

2.8 The Embedded model.

2.9 Four key steps for the feature selection process [40].

3.1 dataset of breast cancer.

3.2 Variation of classification accuracy as a function of the number of selected features during a selected features processing.

3.3 shows Membership values between the examples and two clusters using fuzzy K_means.

Lists of tables:

1.1 Stages of breast cancer [3].

1.2 Treatement of breast cancer depend to stages.

1.3 The statistics of survival after treatment of breast cancer according Canadian Cancer Society 2014. [5].

1.4 The 10 countries most affected by breast cancer [6].

2.1 Some work on the Selection of Variables and classification.

3.1 The experimental result for breast dataset using PSO_SVM algorithm approach.

3.2 Parameters setting.

3.3 The relevant features selected by wrapper filter PSO combinig with SVM.

3.4 k-means and fuzzy k-means clustering for breast data.

Glossary:

WHO: world health organization.

WCR: world cancer research.

ANN: artificial neural networks.

DT : decision trees.

LDA: Linear discriminants analyze.

KNN: k nearest neighbor.

SVM: support vector machine.

FCM: fuzzy c means.

SOM: self organization maps.

PSO: particle swarm optimization.

ACC: accuracy.

AUC : area under the cuve.

Introduction:

Technological advances have facilitated the acquisition and collection of many data, in particular in the medical field during patient examination. These data can be used as medical decision support, leading to development of tools as well as a good understanding of the biological mechanisms involved in breast cancer progression. In the literature, we regularly find the concept of aid in the diagnosis, these systems are even considered essential in many disciplines, these systems based on techniques from artificial intelligence problems but the most interesting are often based on high-dimensional data. These problems denote the situations where we have few observations while number of explanatory variables is very large. This situation is increasingly common in applications, especially those related to biochips.

A biochip provides a single observation of several thousand genes simultaneously. This observation generally corresponds to a single experimental condition and one class (healthy or cancerous cell). Genes play the role of variables, and the number of samples of biochips is very low for reasons cost.

This work is in the context of medical aid diagnosis, which aims to reduce the number of variables among which are low informative and others are essentially noise.

The manuscript is organized as follows:

The first chapter presents the background related to breast cancer and problematic faced in this area of the data dimension.

The second chapter reviews the state-of-the-art of machine learning in cancer research. We have described the three machine learning tasks mostly used in cancer management: supervised classification, clustering and feature selection. A few examples of the most known approaches for each task are briefly described by highlighting their advantages and drawbacks. Then some application examples of such approaches in breast cancer management are provided. This chapter ends with a description of the recent challenges that have to be faced to improve breast cancer management.

The third chapter addresses the problem of data dimensionality by taking advantage of learning capabilities. We particularly propose a wrapper feature selection approach using partical swarm optimization (PSO) combinig with support vector machines (SVM) and these results have implemented in either supervised classification we have used support vector machine learning or unsupervised classification using k-means algorithm suit on a fuzzy rule-based.

Finally, a general conclusion and perspectives of this work Master will be presented.

1. Background:

1.1 Introduction:

Relevant feature identification has become an essential task to apply data mining algorithms effectively in real world scenarios. Therefore, many feature selection methods have been proposed it is the process of detecting relevant features and removing irrelevant ,redundant or noisy in literature to achieve their objectives of classification and clustering to improve inductive learning either in term of generalization capacity learning speed or reducing the complexity of induced model is used in several medical application such as the prediction of efficiency medical tests ,classification of tumors ,cancer detection .

As a result ,this topic is one of the hottest research in bioinformatics ,because the cancer is one of the most common causes of death in world , the most common cancers in terms of incidence were : lung {1.52 million cases}, breast {1.29 million cases} and colorectal {1.15 million cases}are the greatest scourges in humanity .

We focus in our work on breast cancer as one of most frequently diagnosed malignancy in women in the world, according to the world health organization (WHO) is estimated that worldwide over 508000 women died in 2011 due to breast cancer almost 50% of breast cancer cases and 58% of death occur in less developed countries such as Africa.

1.2 Impact of breast cancer:

According to world cancer research (WCR) 26.4million people per year may be diagnosed with cancer by 2030, with 17 million people dying from it. In 2014, the incidence of mortality of the most common cancers in ALGERIA shows that breast cancer can occur in both men and women but it's far more common in women, for that has the first position in women [1] (Figure 1.1).

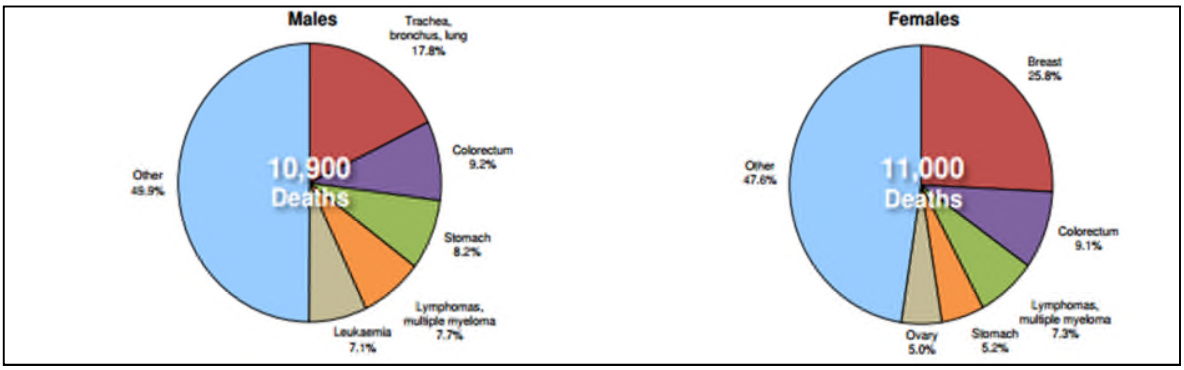

Figure 1.1: Distribution of mortality of the most common cancers in ALGERIA according to the World Health Organization in 2014.

We show in Figure 1.2, breast cancer positions for female gender, the incidence estimates of the most common cancers according to the record of World Health Organization in 2014.

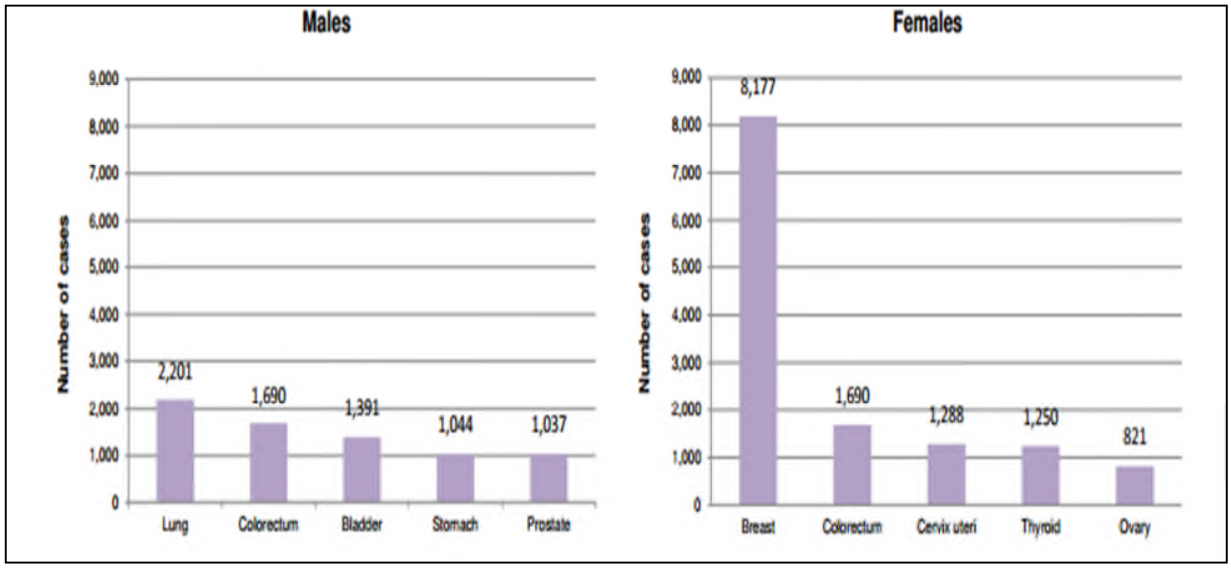

Figure 1.2: the incidence of the most common cancers according to the central register World Health Organization in 2014.

1.3 Definition of breast cancer:

To understand breast cancer, it helps to have some basic knowledge about the normal structure of the breasts, shown in (Figure1.3) The female breast is made up mainly of lobules (milk-producing glands), ducts (tiny tubes that carry the milk from the lobules to the nipple), and stroma (fatty tissue and connective tissue surrounding the ducts and lobules, blood vessels, and lymphatic vessels).

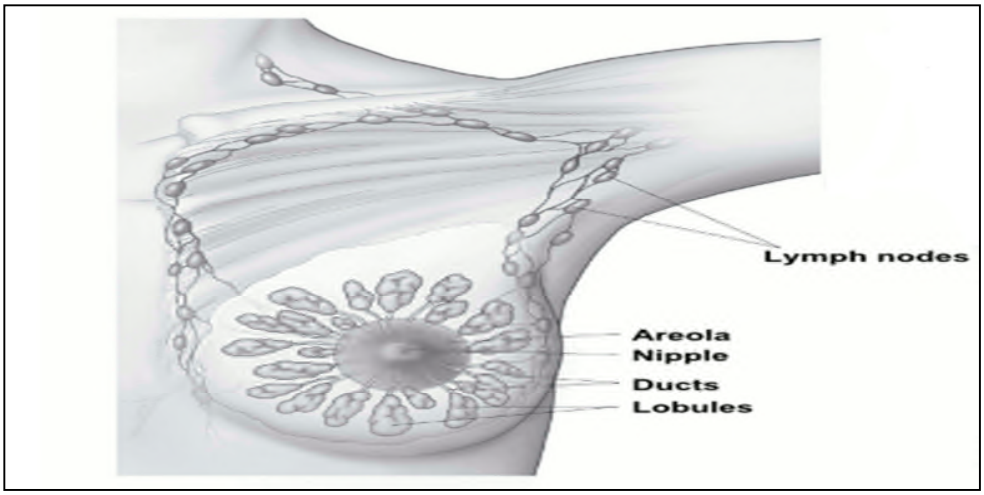

Figure 1.3: Normal breast tissue [2].

Breast cancer is uncontrolled growth of breast calls, it refers to malignant tumor. Most breast cancers begin in the cells that line the ducts (ductal cancers 85 - 90% of all cases), some begin in the cells that line the lobules (lobular cancers 8% of all cases), while a small number start in other tissues.

1.4 Symptoms of breast cancer:

Signs and symptoms of breast cancer may include:

- A breast lump or thickening that feels different from the surrounding tissue.
- Bloody discharge from the nipple.
- Change in the size, shape or appearance of a breast.
- Changes to the skin over the breast, such as dimpling.
- A newly inverted nipple.
- Peeling, scaling or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin.
- Redness or pitting of the skin over your breast, like the skin of an orange.

1.5 Risk factors of breast cancer:

- Gender: Simply being a woman is the main risk factor for developing breast cancer. Men can develop breast cancer, but this disease is about 100 times more common among women than men.
- Aging: the risk of developing breast cancer increases as you get older. About 1 out of 8 invasive breast cancers are found in women younger than 45, while about 2 of 3 invasive breast cancers are found in women age 55 or older.
- Genetic risk factors: About 5% to 10% of breast cancer cases are thought to be hereditary as:
  - BRCA1 and BRCA2: The most common cause of hereditary breast cancer is an inherited mutation in the genes. (The signification of these genes look at Annex A)
  - Changes in other genes: Other gene mutations can also lead to inherited breast cancers such as ATM -P53-CHEK2-CDH1-STK11-PALB2. (the signification of these genes look at Annex A)
- Family history: Breast cancer risk is higher among women whose close blood relatives have this disease.

- Personal history: A woman with cancer in one breast has a 3- to 4-fold increased risk of developing a new cancer in the other breast or in another part of the same breast.
- Dense breast tissue: when they have more glandular and fibrous tissue and less fatty tissue. Women with dense have a risk of breast cancer that is 1.2 to 2 times that of women with average breast density.

1.6 ***Stages of breast cancer:*** Cancers are generally classified in stages according to the size of tumors, and especially the degree of the disease spreading to other parts of the body (see Table 1.1)

stage	Definition
Stage 0	-Cancer cells remain inside the breast duct, without invasion into normal adjacent breast tissue.
Stage IA	-The tumor measures up to 2 cm AND the cancer has not spread outside the breast, no lymph nodes are involved
Stage IB	-There is no tumor in the breast, instead, small groups of cancer cells -- larger than 0.2 millimeters but not larger than 2 millimeters – are found in the lymph nodes. OR -there is a tumor in the breast that is no larger than 2 centimeters, and there are small groups of cancer cells – larger than 0.2 millimeter but not larger than 2 millimeters – in the lymph nodes.
Stage IIA	-No tumor can be found in the breast, but cancer cells are found in the axillary lymph nodes (the lymph nodes under the arm) OR -the tumor measures 2 centimeters or smaller and has spread to the axillary lymph nodes. OR -the tumor is larger than 2 but no larger than 5 centimeters and has not spread to the axillary lymph nodes.
Stage IIB	-The tumor is larger than 2 but no larger than 5 centimeters and has spread to the axillary lymph nodes. OR -the tumor is larger than 5 centimeters but has not spread to the axillary lymph nodes.
Stage IIIA	-No tumor is found in the breast. Cancer is found in axillary lymph nodes that are sticking together or to other structures, or cancer may be found in lymph nodes near the breastbone . OR -the tumor is any size. Cancer has spread to the axillary lymph nodes, which are sticking together or to other structures, or cancer may be found in lymph nodes near the breastbone.
Stage IIIB	-The tumor may be any size and has spread to the chest wall and/or skin of the breast .AND -may have spread to axillary lymph nodes that are clumped together or sticking to other structures or cancer may have spread to lymph nodes near the breastbone. Inflammatory breast cancer is considered at least stage IIIB.
Stage IIIC	-There may either be no sign of cancer in the breast or a tumor may be any size and may have spread to the chest wall and/or the skin of the breast AND -the cancer has spread to lymph nodes either above or below the collarbone AND -the cancer may have spread to axillary lymph nodes or to lymph nodes near the breastbone.
Stage IV	-The cancer has spread — or metastasized — to other parts of the body.

Table 1.1: stages of breast cancer [3].

1.7 Treatement: [4]

The proposed treatments for Breast cancer depends largely on the stage of disease at diagnosis (Table 1.2) and Treatment is not always necessary for stage 0 breast cancer. Other factors such as the location of the tumor or the patient's general health status also have their importance and will adapt treatments to their situation.

Stage	Treatments
Stage I	-Local therapy: can be treated with either breast conserving surgery (BCS sometimes called lumpectomy or partial mastectomy) or mastectomy + Radiation therapy .
Stage II	- Local therapy: Stage II cancers are treated with surgery . - Systemic therapy: is recommended for women with stage II breast cancer. It may be hormone therapy, chemo, HER2 targeted drugs .
Stage III	- These cancers are treated with chemo before surgery (neoadjuvant chemo), and often radiation therapy is needed after surgery.
Stage IV	- Although surgery and/or radiation may be useful in some situations, systemic therapy is the main treatment. Depending on many factors, this may consist of hormone therapy, chemotherapy, targeted therapies, or some combination of these treatments.

Table 1.2: treatement of breast cancer depend to stages.

After the different treatments that have been located by their stages a rate survival in 2014 was done for Canadian Cancer Society (Table 1.3).

Stage	Relative survival rate after 5 years
0	100 %
I	100 %
II	86 %
III	57 %
IV	20 %

Table 1.3: The statistics of survival after treatment of breast cancer according Canadian Cancer Society 2014. [5]

Breast cancer to the Worldwide according to the World Cancer Research, which estimates the 10 countries most affected by breast cancer over 100,000 people (Table 1.4).

Country	Age-Standardised Rate per 100,000 (World)
Belgium	111.9
Denmark	105.0
France (metropolitan)	104.5
The Netherlands	99.0
Bahamas	98.9
Iceland	96.3
United Kingdom	95.0
Barbados	94.7
United States of America	92.9

Table 1.4: the 10 countries most affected by breast cancer [6].

2. Diagnostic support:

Today the difficulty lies not only in obtaining genomic and proteomic data but also in their analyzes, the objective is to develop analysis methods to extract maximum information from the data collected by biologists and geneticists, they made a big emerge number of problems, it is clear that a good selection procedure in practice be completely explicit, easy to implement and easy to calculate.

So that The selection of biological data contribute to the increased medical diagnosis, the level and growth rate of biomarkers measured repetitive manner on each subject to quantifier the severity of the disease and susceptibility to grow, This is usually interesting to clinical and scientific area.

For the reason that help the expert to make these decisions in a less last time than the survival of a patient. The typical field of such a situation is the biomedical field where we can now do a lot of measurements on a given person for example measuring gene expression also in the case study of a disease, the number of carriers of the disease who participate in study is often limited. The area that concerns the development of methods which allow selection of relevant variables is very active, can provide better prediction and correctly select these variables is important for the classification of disease and the interpretation of the model using fuzzy logic because a cliniticien

will obviously be interested to know that such and such genes are involved in the development of metastasis.

3. Problematic:

The selection of variables has become the object that attracts the attention of many Researchers in recent years, this selection allows identifier and eliminates variables that penalize the performance of a complex model. In addition, the identification of the relevant variables facilitate interpretation and comprehension medical and biological aspects, thus it improves performance prediction methods classification and override the high dimensionality of the data.

The specific problem as the variable selection requires an approach particular since the number of variables is far higher the number of experiments or observations in the literature machine Learning approaches, this approach is to browse the selection of variables before process of learning and keeps only the informative features.

The work presented in this study is in selection of particular variables breast cancer genes that allows the context of developing diagnostic support for detecting the patient (unhealthy or healthy) and could bring more knowledge about the characteristics of this cancer. Also, we highlight the use of selected informative features in classification with support vector machine and using K_means as clustering algorithm and combinig with fuzzy logic to better understand the domain and improve classification rates for decision support.

4. Conclusion:

The selection and classification with fuzzy and non fuzzy approch of the biological data is a field which has been the subject of several researches, as they are described in Chapter 2 feature selection and classification contribute to the strengthening of medical diagnostic by intelligent recognition of biological data.

1. Introduction:

For a long time cancer management was performed based on expert qualitative knowledge held by individuals or using diverse medical guidelines. However, cancer disease has been shown to be complex and very heterogeneous which make the qualitative approach insufficient and the decision-making process very complicated. Breast cancer diagnosis for instance is based on the analysis of thousands of mammograms issued by imaging detection tools. This important task seems to be very complex and tiring and can even lead the radiologists to commit some diagnosis errors. Furthermore, the prognosis task involves usually multiple physicians with different skills using different biomarkers and clinical factors. Typically in such cases many types of qualitative information are integrated to come up with a reasonable decision about the prognosis by the attending physicians based on their own intuition. This is not an easy task even for the most skilled clinicians. If we add to that the increased need to explore the large amount of biological data being available (proteomic and genomic measurements) more efficient approaches to help physicians in their day-to-day practices have become indispensable. Recently, machine learning has been shown very effective to help physicians in their decision making by constructing more accurate prediction and classification models.

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy, heterogeneous or complex datasets.

Generally, machine learning methods are used to analyze medical datasets organized in table form containing a set of patient (individuals or patterns) in term of their properties (attributes, features, variables). The use of machine learning methods in cancer research can be summarized in three main tasks:

- Classifying new patients based on trained models to already-defined cancer classes, known as supervised classification.
- Regrouping patients having similar properties into subgroups, known as unsupervised classification or clustering.
- Selecting relevant biomarkers using feature selection approaches either in a supervised or unsupervised context.

2. **Supervised classification:**

Classification is considered as one of the fundamental problems in machine learning. Duda and Hart (2001) define it as the problem of assigning an element or instance to one of several pre-specified categories. Only available information is a set of patterns characterized by a set of features each of them assigned to a predefined class. Each pattern is classified based on a set of classification rules which are often unknown in many real-life situations. As a simple example, we can cite the problem of breast cancer diagnosis as a supervised classification problem (Wolberg et al., 1994). Almost all machine learning approaches applied on this problem employ supervised Learning such as artificial neural networks (Rumelhart et al., 1986), decision trees (Quinlan, 1986), discriminant analysis (Fisher, 1936), k- nearest neighbor (Cover and Hart, 1967) and Support Vector Machines (Vapnik, 1998). We list below some of the most used supervised machine learning approaches in cancer research.

2.1 Artificial neural networks:

Artificial neural networks (ANN) were originally inspired from the human-being brain which works with interconnected neurons (Figure 2.1). The strength of neural connection is determined through a learning process on labeled data characterized by weights. In an ANNs, the neurons are organized in layers, in such a way that usually only neurons belonging to two consecutive layers are connected.

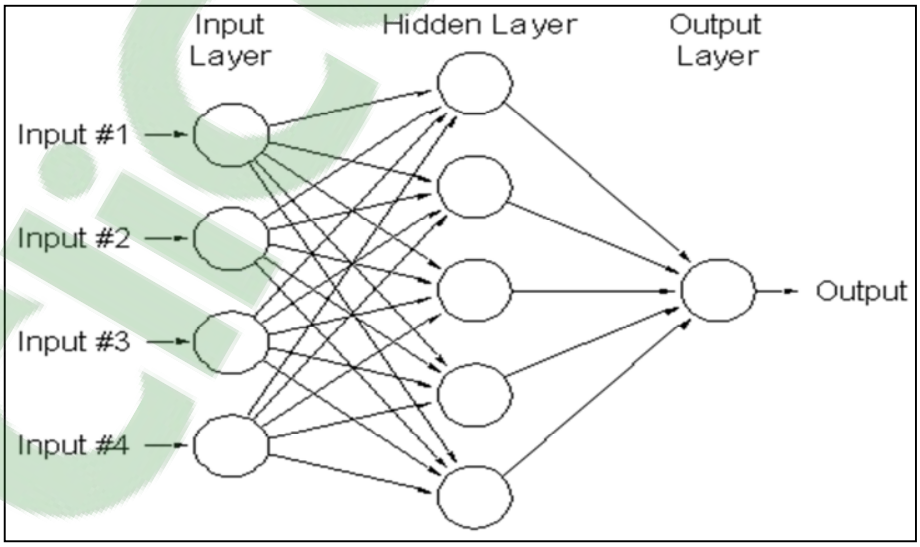

Figure 2.1: Artificial Neural Network [7].

2.2 Decision trees:

A decision tree (DT) is a graphical representation or flow chart of decisions (nodes) and their possible consequences (leaves or branches) used to create a plan to reach a goal (Quinlan, 1986) (Figure 2.2). In a classification tree, pattern classification starts from the root node by successively asking questions about each of its properties (features). Different exclusive links from a root node correspond to the different possible values of the property (feature).
According to the answer, this process is followed until arriving to a leaf node which has no further question. The pattern is finally assigned to the class represented by this node.

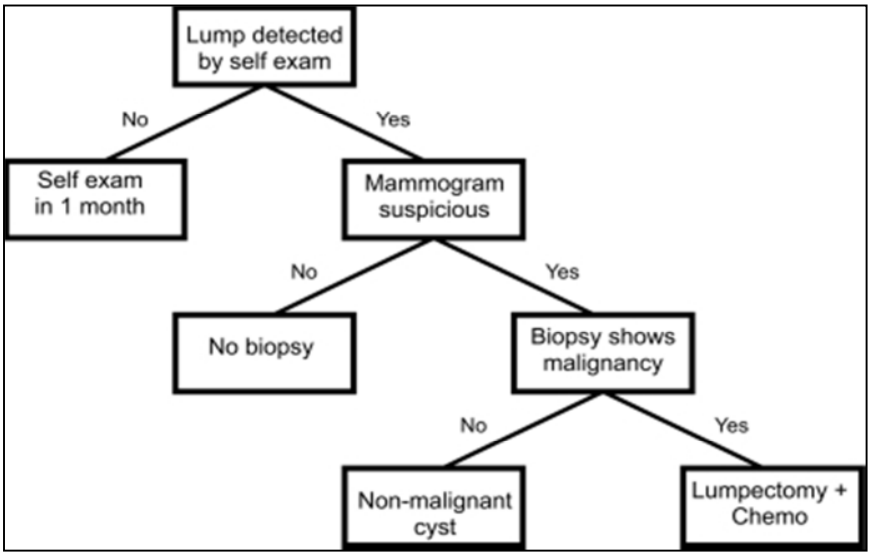

Figure 2.2: Decision tree that might be used in breast cancer diagnosis and treatment [8].

A variety of approaches can be found for choosing the appropriate order of features in the decision tree and how possibly make reduce the large trees. Decision trees are very well accepted in medical applications owing to its high model transparency, comprehensive and interpretability. Some potential limitations affecting the application of decision trees in cancer research is its difficulty to scale with high dimensional data (e.g. microarray data) and the strong assumption on mutual exclusivity of classes [9].

2.3 Discriminant analysis:

Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant analysis (Fisher, 1936) constructs a linear hyperplan based on the maximization of between-group to within-group ratio. Assuming a multivariate normal distribution and homogeneity of covariance matrices, the hyperplan is described by a linear discriminant function which equals zero at the hyperplan. In this case, the hyperplan is defined by geometric means between the centroids (i.e. the center of each classe).Recently, a variety of non linear discriminant analysis approaches were proposed based on kernel concept to improve its classification performance. This approach has found its place in some breast cancer applications ([10]; [11]; [12]; [13]).However, this approach suffers from several limitations such as the small sample size problem due to within-class matrix singularity [14]. This problem arises whenever the number of samples is smaller than the dimensionality of samples (the case of cancer classification with gene expression profiling characterized by thousands of genes and less than one hundred patients).

2.4 k- nearest neighbor:

The k- nearest neighbor method classifies each unlabelled sample by the majority label among its k nearest neighbors in the training set (Cover and Hart, 1967). This makes it very well suited for non-linear classification problems. One potential of this approach is that it does not make any assumption on data distribution. A variety of breast cancer studies can be found in literature based on this approach ([15]; [16]; [17]). Though simple however, it is known that k-NN classifier is very sensitive to the presence of irrelevant features. Moreover, this method tends to be slow for large training dataset because the nearest neighbors should be searched over all instances [18].

2.5 Support vector machines:

The key idea of this approach is that by an appropriate mapping into sufficiently high dimensional space, it is always possible to define a hyperplan that separates the data from two categories (Vapnik, 1998) (Figure 2.3).

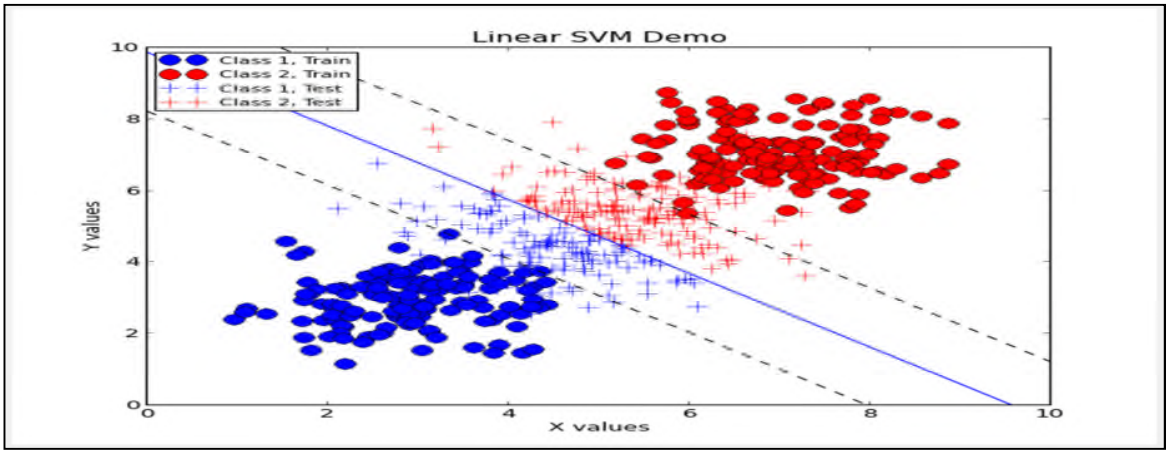

Figure 2.3: Support Vector Machines [19].

The mapping is performed using some specific functions (known as kernel functions) which are chosen by the user among a variety of functions (Gaussian, polynomial, linear) according to the problem under investigation. The goal in all cases is to find the separating hyperplan in the resulted space with the largest margin, expecting that the larger is the margin, the better is the generalization of classifier (Vapnik,1998). This problem is generally reformulated as a constrained optimization problem and solved generally by resorting to its dual reformulation. Various applications using SVM has been performed on breast cancer research ([20]; [21]; [22]).

3. **Unsupervised classification (clustering):**

Clustering is considered as one of the fundamental research problems in various data analysis fields such as machine learning and pattern recognition. Cluster analysis focus to organize a set of patterns (e.g. patients or genes) into clusters such that patterns within a given cluster have a high degree of similarity, whereas patterns belonging to different clusters have a high degree of dissimilarity. Unlike supervised classification, the outcome of each element in the unsupervised context is unknown making the learning task more challenging. One typical example in cancer research is the clustering of genes expression data. In microarray experiment, the expression value of thousands of genes is

obtained for only few patients. Extracting co-expressed genes in different samples from this data is of great importance as it may allow gaining new insights into cancer biology. This is typically a clustering problem where co-expressed genes should be grouped into the same cluster.

Many algorithms have been proposed to address this problem for different purposes. Clustering techniques can be roughly divided into two main categories: Hierarchical and partitioning.

3.1 Hierarchical clustering:

Hierarchical clustering produces a nested series of partitions on the form of tree diagram or dendogram. In hierarchical clustering we can distinguish two situations between two groups from different partitions: either they are disjoint or one group wholly contains the other (Figure 2.4). Two clusters are merged in hierarchical measure based on a distance or dissimilarity measure such as Minkowski and Mahalanobis measures. It exist several algorithms to establish a hierarchical tree: agglomerative and divisive.

Hierarchical clustering is the most commonly used method to summarize data structures in bioinformatics generally and in breast cancer specifically [18]. Many studies can be found in cancer research literature about the use of this clustering approach, especially for microarray data analysis.

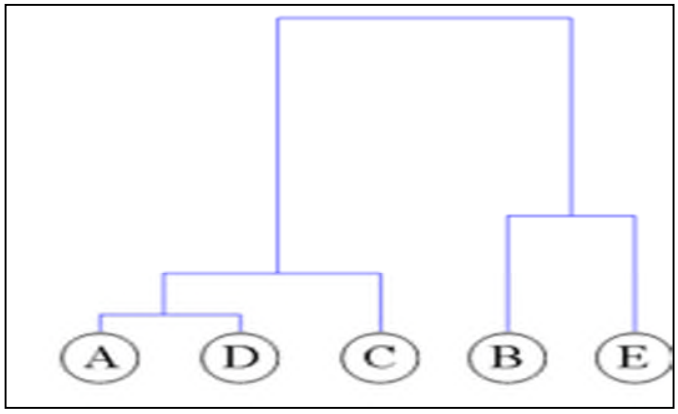

Figure 2.4: Hierarchical clustering [23].

3.2 Partitioning clustering:

Partitioning clustering identifies only one partition of the data that optimizes an appropriate objective function (kernel, spectral, fuzzy and classical) (Figure 2.5).

The clustering can be either hard (each pattern belongs to only one class) or fuzzy where each pattern belongs with a certain degree of membership to each resulting cluster. Fuzzy clustering offers the advantage to provide a basis for constructing rule-based fuzzy model that has simple representation and good performance for non-linear problems.

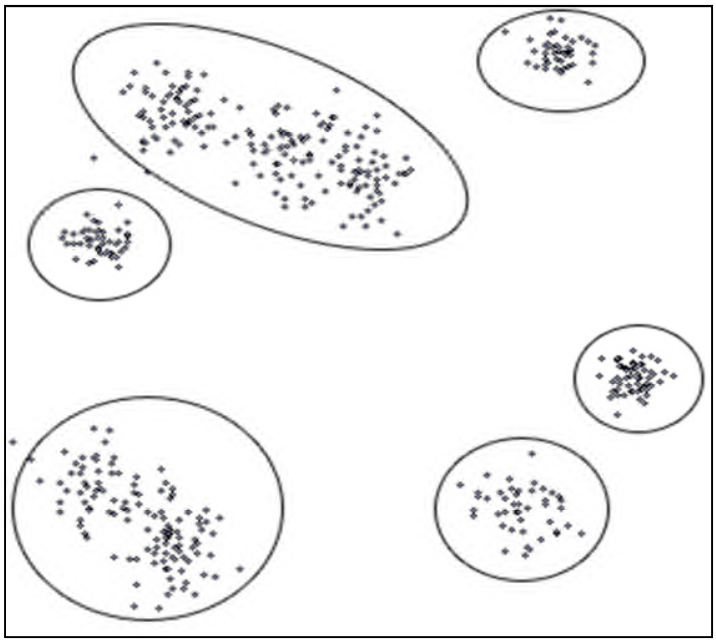

Figure 2.5: Partitioning clustering [24].

The k-means algorithm (MacQueen, 1967) is one of the most popular partitioning clustering algorithms. This algorithm is based on a "hard" partition of the data into k clusters based on the minimization of the within-group sum of squares. A direct extension of the k-means algorithm is the Fuzzy C-means (FCM) (Bezdec, 1981), where the fuzzy set notion is introduced into the class definition in this case each element belongs to a given class with certain membership degree. These clustering approaches are widely used in breast cancer research. For instance a molecular classification of tumor samples can be achieved using either unsupervised methods like k-means clustering ([26];[27]) or 'SOMs' (self organizing maps) [28].

Clustering approaches have been also used to cluster the gene in groups and establish the relation between the coexpressed genes in each group [29].

Many studies can be found also where the clustering is performed in both directions, i.e. patients and genes, called biclustering. However, the use of different methods may yield different results.

4. Feature selection:

The features selection is a very uses in the context of data Manning in this recent year. it is crucial for applications of very large base such as genetic engineering, complex industrial processes. This is actually summarizing and intelligently extracts knowledge from raw data. The value of variable selection is summarized in the points following:

-When the number of variables is really too large, the learning algorithm cannot complete execution within a reasonable time, then selection can reduce the feature space.

-From an intelligence perspective artificial create a classifieur is to create a model for the data. Or a legitimate expectation for a model is to be as simple as possible.

- It improves the performance of the classification: its speed and power generalization.
- It increases the comprehensibility of data.

This process focus on selecting relevant features, there are three general approaches for feature selection the literature:

- Wrapper Approach.
- Filter Approach.
- Embedded Approach.

4.1 Wrapper Approach:

The wrappers were introduced by John and al. in 1994 [30]. Their strategies for feature selection use an induction algorithm to estimate the merit of feature subsets. The rationale for wrapper approaches is that the induction method that will ultimately use the feature subset should provide a better estimate of accuracy than a separate measure that has an entirely different inductive bias [31] .Feature wrappers often achieve better results than filters due to the fact that they are tuned to the specific interaction between an induction algorithm and its training data. However, they tend to be much slower than feature filters because they must repeatedly call the induction algorithm and must be re-

run when a different induction algorithm is used. Since the wrapper is a well defined process, most of the variations in its application are due to the method used to estimate the off-sample accuracy of a target induction algorithm, the target induction algorithm itself, and the organization of the search. The graphical representation of the wrapper model is shown in (Figure 2.6).

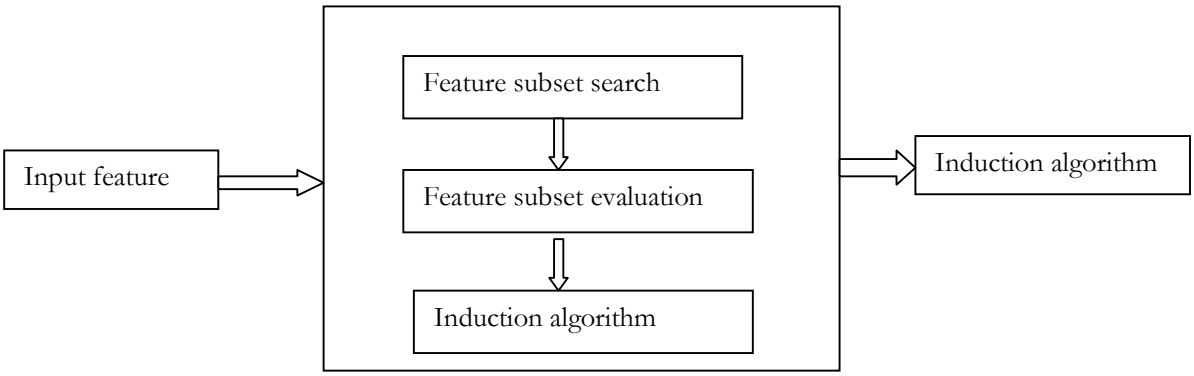

Figure 2.6: The wrapper model [32].

4.2 Filter Approach:

The earliest approaches to feature selection within machine learning were filter methods. All filter methods use heuristics exploits the general characteristics of training data with independent of the mining algorithm to evaluate the merit of feature subsets. As a consequence, filter methods are generally much faster than wrapper methods. One of these advantages is to be completely independent of the data model that we seek to build. It offers a subset of variables for satisfying explain the structure of data hiding and that subset is independent the learning algorithm selected. This context is also adaptive in selection unsupervised variables ([33]; [34]; [35]). Furthermore filter approch are generally less costly in computation time since they avoid repetitive executions learning algorithms on different subset of variables. In contrast, their major disadvantage is that they ignore the impact of selected sets in the performance of the learning algorithm. The graphical representation of the filter model is shown in (Figure 2.7).

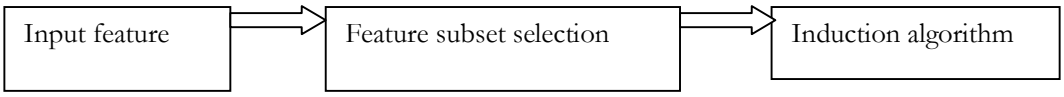

Figure 2.7: The Filter model [32].

4.3 *Embedded Approach:*

This approach interacts with learning algorithm at a lower computational cost than the wrapper approach. It also captures feature dependencies. It considers not only relations between one input features and the output feature, but also searches locally for features that allow better local discrimination. It uses the independent criteria to decide the optimal subsets for a known cardinality. And then, the learning algorithm is used to select the final optimal subset among the optimal subsets across different cardinality. The graphical representation of the embedded model is shown in (Figure 2.8).

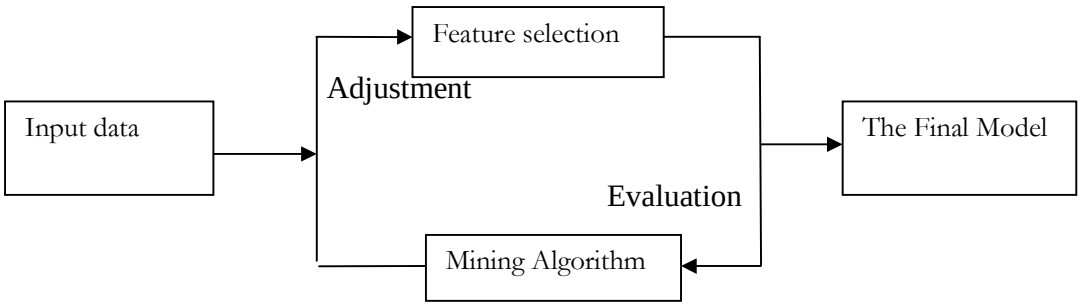

Figure 2.8: The Embedded model.

Besides, the general procedure for feature selection has four key steps as shown in (Figure 2.9).

- Subset Generation
- Evaluation of Subset
- Stopping Criteria
- Result Validation

Subset generation is a heuristic search in which each state specifies a candidate subset for evaluation in the search space. Two basic issues determine the nature of the subset generation process. First, successor generation decides the search starting point, which influences the search direction. To decide the search starting points at each state, forward, backward, compound, weighting, and random methods may be considered [36]. Second, search organization is responsible for the feature selection process with a specific strategy, such as sequential search, exponential search [37] or random search [38]. A newly generated subset must be evaluated by a certain evaluation criteria. Therefore, many evaluation criteria have been proposed in the literature to determine the goodness of the candidate

subset of the features. Base on their dependency on mining algorithms, evaluation criteria can be categorized into groups: independent and dependent criteria [39]. Independent criteria exploit the essential characteristics of the training data without involving any mining algorithms to evaluate the goodness of a feature. And dependent criteria involve predetermined mining algorithms for feature selection to select features based on the performance of the mining algorithm applied to the selected subset of features. Finally, to stop the selection process, stop criteria must be determined. Feature selection process stops at validation procedure. It is not the part of feature selection process, but feature selection method must be validate by carrying out different tests and comparisons with previously established results or comparison with the results of competing methods using artificial datasets or real world datasets or both.

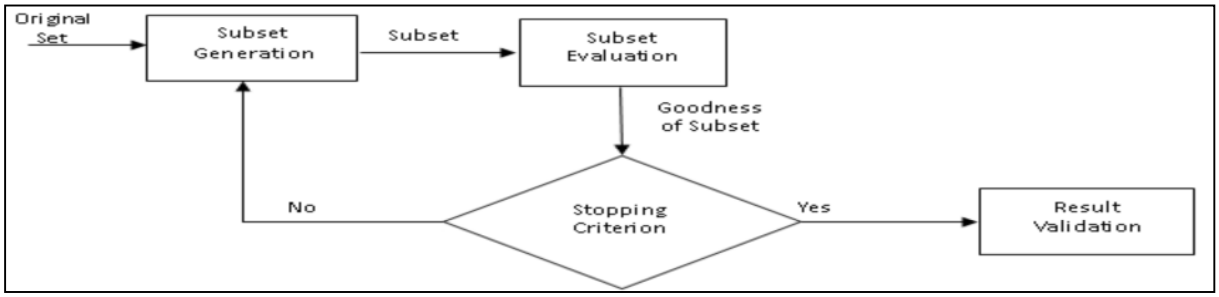

Figure 2.9: Four key steps for the feature selection process. [40]

5. Selections of feature and classification in literature:

In this table we present some work on the resolution of the problem with a very large base also we present methods and their applications in different areas so the combining of variable selection in learning and classification in one step. And some state of the art work that contributed to the selection of breast cancer. (Table 2.1).

Authors	**Articles**	**Approaches**	**Experiences**	**Results**
Yuhang Wing, Fillia Ma kedon, 2004 [41]	Application of ReliefF to selecting informative genes for cancer classification using microarray data.	This paper implements the selection method ReliefF for select genes the most relevant of different basic data with the SVM classifier and K-NN.	The basics data: ALL leukemia, MLL leukemia.	After selection 150 genes for each basis, classification rates are: SVM: - ALL: 99% - MLL:97% K-NN: - ALL: 100% - MLL: 98%
TianLan, Deniz Erdogmus Andre, Adami Michael Pavel, 2005 [42]	Feature selection by ICA and MIM in EEG Signal Classification.	This paper proposes a selection scheme variable using linear analysis independent component and the mutual information principle is to maximize Information. evaluation rate classification was made with the classifieur K-nn.	The experiments of this article were produced by the use of EEG signal.	Several tests were made with different number of variables selected from: 20 variable rate is 82%, 30 variables a rate is 87%, After selecting 35 variables are noted a fall the rate classification with 2%.
Shousken Li, Rui Xia, Chingquing Zong, Chui Ran Hueing, 2009. [43]	A framework of feature selection methods for text categorization.	This paper focuses on the classification of texts, is based on the selection of terms and their classification, compare six methods: DF (document frequency), MI (mutual information, IG (information gain), CHI-2 (X2 test, SNB (bi-normal separation) and WLLR	experiences were tested on a body of Reters-21578 called R2 and 20ng is a collection of about 20000 Under 20 documents.	DFscore =0,004 MIscore =0.870 Which shows that MI score expressed a good information The category.

		(weighted loglikelihood ratio), these methods have been implemented to measure the score between the terms and their categories.		
Mehmet Fatih Akay * 2009 [44]	Support vector machines combined with feature selection for breast cancer diagnosis	This paper focuses on the Classification the Wisconsin Diagnostic Breast Cancer (WDBC) data set Using SVM with feature selection using F-score	Use of the Wisconsin Diagnostic Breast Cancer (WDBC) data set with the 32 original features for the training phase.	the F-score + SVM reduces the input scale by transforming the original data into a new format of 5 features and the accuracy is 99.51%
Prasad, Y.,Al 2010 [45]	Svm classifier based feature selection using ga, aco and pso for sirna design	This paper focuses on the Classification the Wisconsin Diagnostic Breast Cancer (WDBC) data set Use of support vector machine and ga, aco and pso algorithms	Use of the Wisconsin Diagnostic Breast Cancer (WDBC) data set with the 32 original features for the training phase	After selection And classification for each algorithm obtained : - ACO-SVM : 15 features and the accuracy is 95.96% - GA-SVM : 18 features and the accuracy is 97.19% - PSO-SVM : 17 features and the accuracy is 97.37%
Mohammad Javad Abdi,SeyedMohammd Hosseini, and Mansoor Rezghi 2012 [46]	A NovelWeighted Support Vector Machine Based on Particle Swarm Optimization for Gene Selection and Tumor Classification	This paper mRMR is applied to filter out many unimportant genes and reduces the computational load for SVM classifier. Then, a PSO-based approach is developed for determination of kernel parameters and genes weight.	The basics genetic data: genes= 7129 ALL leukemia, MLL leukemia. And genes= 2000 Colon	After selection And classification for each basis use mRMR-PSO-WSVM are : -leukemia data: accuracy 100 % gene selected =3.8 -colon data: accuracy 93.55% gene selected = 6.2
Bichen Zheng, Sang Won Yoon ⇑, Sarah S. Lam	Breast cancer diagnosis based on feature	This paper focuses on the Classification the	Use of the Wisconsin Diagnostic	the K-SVM reduces the input scale by transforming the original data into a new format of 6 features

2014 [47]	extraction using a hybrid of K-means and support vector machine algorithms.	Wisconsin Diagnostic Breast Cancer (WDBC) data set Use a hybrid of K-means and support vector machine (K-SVM) algorithms	Breast Cancer (WDBC) data set with the 32 original features for the training phase	and the accuracy is 97.38% and CPU time (seconds) K-SVM 6 0.0039
Khalid Raza 2014 [48]	Clustering analysis of cancerous microarray data.	this paper, we applied four different clustering techniques, such as k-means, hierarchical, density-based and expectation maximization approaches, on five different kinds of cancerous gene expression data (lung, breast, colon, prostate, breast and ovarian cancer) for their analysis	Cancer data in this paper, i.e. breast cancer is taken from published papers [49].	after normalizing data we used a two tailed t-test for extracting differentially expressed genes among two types of sample, i.e., normal and tumor. for breast cancer data the result on the basis of normal and tumor cluster i.e. 1 and 0. The percentage values signifies that number of instances participated or concerned to that : - K_means: 0 -> 58% 1 -> 42% -HC: 0 -> 99% 1-> 1% - Densitybase: 0 -> 28% 1-> 72% - EM : 0 -> 100% 1-> 0%

Table 2.1: Some work on the Selection of Variables and classification.

6. Recent challenges in breast cancer management:

In spite of the intensive research performed in the machine learning filed in past decades, many challenges are still needed to be addressed seriously to improve cancer management. Challenges are mainly related to data characteristics used in decision-making process. Three challenges are mainly faced: the first one is related to the presence of mixed-type data in daily produced clinical datasets, the second one is related to high dimensionality in data especially issued from microarray technology and the last one is the problem of noise and uncertainties associated usually to both data. Addressing efficiently those problems is urgently needed provided that in some cancer applications the three challenges can be even faced simultaneously (e.g. integration of clinical and microarray data to improve breast cancer management ([50]; [51])).

7. Conclusion:

In this chapter we have reviewed the state-of-the-art of machine learning in cancer research. We have described the main three machine learning tasks most used in cancer management: supervised classification, clustering and feature selection. A few examples of the most famous approaches for each task have been briefly described by highlighting their advantages and drawbacks.

Although their successful use in breast cancer management based on traditional clinical factors, we have noticed that most of them fail to deal with the recent challenges brought by the introduction of data issued from advanced technologies. We can for instance mention the problem of overfitting in supervised classification methods due usually to the low feature-to-sample ratio. This requires a resort to feature selection approaches extensively studied and developed to overcome this problem. However that features selection is not only useful for dimension reduction but has made major advancements to gain new insights in cancer biology by using gene expression profiles. Thanks to feature selection approaches a tailored and personalized cancer management is today underway by the derivation of several genetic signatures for different purposes.

This chapter ends with a description of the recent challenges that have to be faced to improve cancer management. We considered mainly the problems of data heterogeneity, high dimensionality, and low signal-to-noise ratio and membership uncertainties.

In next chapter we address the problem of high dimensionality through the development of an wrapper feature selection strategy based on support vector machines optimized by particle swarm optimization for relevant and minimum feature subset for obtaining higher accuracy of ensembles and these results have implemented in either supervised classification we have used support vector machine learning or unsupervised classification using k-means algorithm suit on a fuzzy rule-based.
In order to offer the capability to deliver the process of turning data into knowledge that can be understood by biologists and geneticists.

1. **Introduction:**

Pretreatment methods of data give us a mature technology to solve problems or first challenge is to go beyond current learning to deal with this new range of problem dimension reduction, variable selection methods can make some more robust systems. This strength is expressed as the ability of the pretreatment method to produce relevant features allowing best classification whenever the data is disturbed.

In this chapter we develop our contribution represented in the selection variables and classification (see Chapter 2). For this chapter is organized into two main steps: The first focuses on the methods of selection and the second contains the results of classification with different experiments.

2. **Database:**

This data set produced by West et al. [52] concerns breast cancer. It consists of 49 samples including 25 tumor tissues and 24 tissues healthy or normal (Figure 3.1). The experiments were conducted with DNA chips. Falling expression values for 7,129 human genes with the highest minimum intensities were selected.

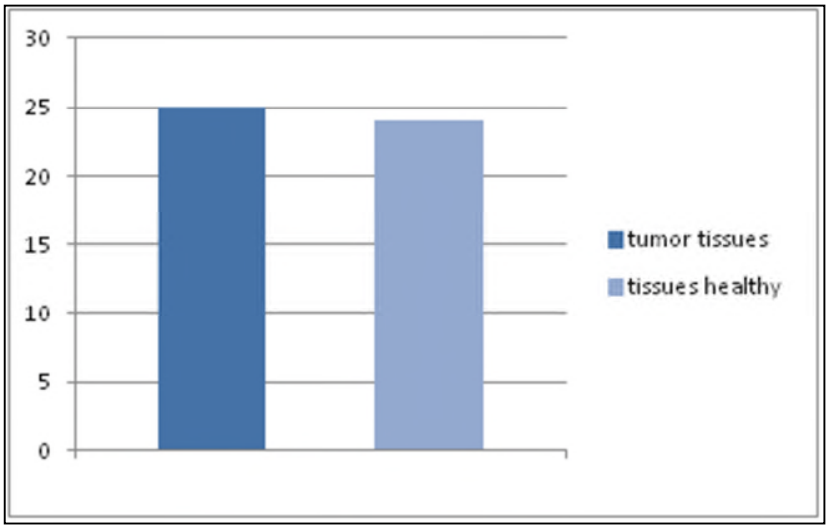

Figure 3.1: dataset of breast cancer.

3. Step of selection:

3.1 *Feature selection using SVM-PSO*

The feature selection is based on the support vector machines (SVM) optimized by particle swarm optimization (PSO) .SVM classifier is a supervised learning algorithm based on statistical learning theory, whose aim is to determine a hyper plane that optimally separates two classes by using train data sets. Assume that a training data set $\{xi,yi\}i=1^n$, where x is the input vector, and y {+1,-1} is class label. This hyper plane is defined as w.x + b = 0, where x is a point lying on the hyper plane, w determines the orientation of the hyper plane and b is the bias of the distance of hyper plane from the origin. The aim is to determine the optimum hyper plane. Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique based on the simulation of the behavior of birds within a flock. The swarm is a population of particles. Each particle represents a potential solution to the problem being solved. The personal best (pbest) of a given particle is the position of the particle that has provided the greatest success (i.e. the maximum value given by the classification method used). The local best (lbest) is the position of the best particle member of the neighborhood of a given particle. The global best (gbest) is the position of the best particle of the entire swarm. The leader is the particle that is used to guide another particle towards better regions of the search space. The velocity is the vector that determines the direction in which a particle needs to "fly" (move), in order to improve its current position. The inertia weight, denoted by W, is employed to control the impact of the previous history of velocities on the current velocity of a given particle. The learning factor represents the attraction that a particle has toward either its own success (C1 -cognitive learning factor) or that of its neighbors (C2 - social learning factor). Both, C1 and C2, are usually defined as constants. Finally, the neighborhood topology determines the set of particles that contribute to the calculation of the lbest value of a given particle.

we describe the hybrid PSOSVM approach for gene selection and classification of Microarray data. The PSO algorithm is designed for obtaining gene subsets as solutions in order to reduce the high number of genes to be later classified. The SVM classifier is used whenever the fitness evaluation of a tentative gene subset is required. (The process of PSO algorithm for solving optimization problems is mentioned look at Annex B).

Our implementation was carried out on the Matlab 8.1development environment by extending the LIBSVM which is originally designed by Chang and Lin [54].

The experimental result for breast dataset using PSO_SVM algorithm approach.(see the table3.1).

Fold#	**PSO features selection SVM**	
	accuracy [%]	**selected features**
1	82.09	9
2	88.06	13
3	85.07	12
4	86.57	18
5	80.60	20
6	88.06	10
7	80.60	19
8	89.55	11
9	80.60	15
10	85.80	14

Table 3.1: The experimental result for breast dataset using PSO_SVM algorithm approach

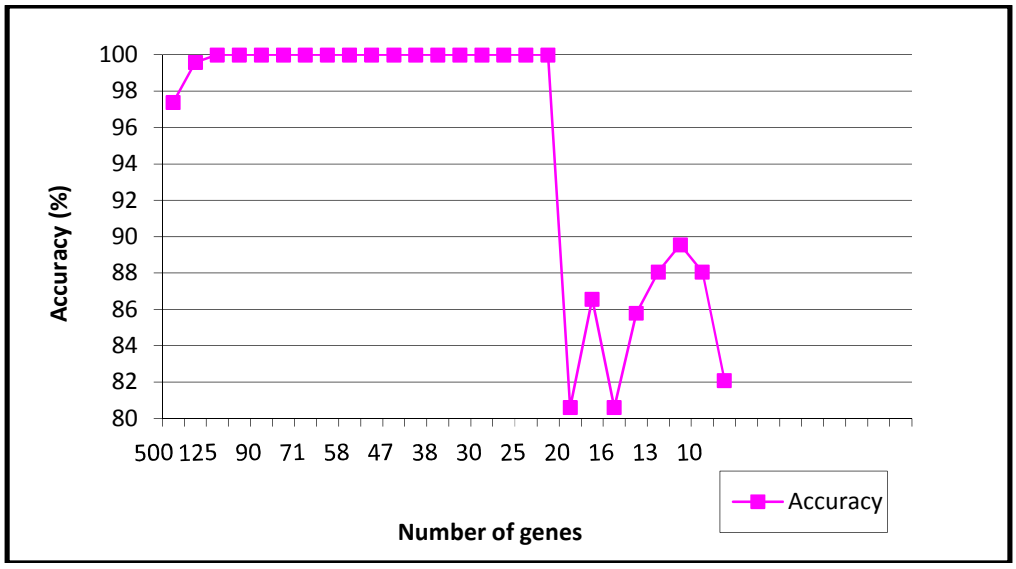

Figure 3.2: Variation of classification accuracy as a function of the number of selected features during a selected features processing.

In evolutionary feature selection as Partical swarm optimisation (PSO) the parameter setting that is able to contribut a hight impact on the classifier‘s performance instead of the size of features for that the termination criterion is that the generation number reaches generation 100. The best indexes of features are obtained when the termination criteria satisfy. The detail parameter setting for partical swarm optimization algorithm is as the following (See the table 3.2).

Parameters	**value**
Population size	20
Number of generations	100
C1, C2	2
Vmax	2

Table 3.2: Parameters setting.

So the application of partical swarm optimization (PSO) with support vector machines (SVM) approach to breast cancer has given a list of genes, we present Brief Biological Analysis of Selected Genes about eleven of the best subsets of genes found. (See the table 3.3).

Gene	**Biological Description [55]**
X58072	GATA3 (protein 3) Human hGATA3 mRNA for trans-acting T-cell specific transcription factor and plays an important role in endothelial cell biology.
X87212	CTSC (cathepsin C) H.sapiens mRNA for cathepsin C.the protein encoded by this gene, a member of the peptidase C1 family, is a lysosomal cysteine proteinase that appears to be a central coordinator for activation of many serine proteinases in immune/inflammatory cells.
L17131	Human high mobility group protein (HMG-I(Y)) gene exons 1-8, complete cds.
X03635	ESR1 (estrogen receptor 1) Human mRNA for oestrogen receptor. Estrogen receptors are also involved in pathological processes including breast cancer, endometrial cancer, and osteoporosis.

M23263	AR (androgen receptor) Human androgen receptor mRNA, complete cds. Expansion of the polyglutamine tract causes spinal bulbar muscular atrophy (Kennedy disease).
HG4716-HT5158	Guanosine 5' -Monophosphate Synthase.
X76180	SCNN1A sodium channel, non voltage gated 1 alpha subunit ,H.sapiens mRNA for lung amiloride sensitive Na+ channel proteinmineralocorticoids.
M29877	FUCA1 , Human alpha-L-fucosidase, complete cds .
HG3494-HT3688	Nuclear Factor Nf-Il6.
U32907	LRRC17 leucine rich repeat containing 17, Human p37NB mRNA, complete cds.
D79206	SDC4 syndecan 4, Human gene for ryudocan core protein, exon1-5, complete cds.

Table3.3: the relevant features selected by wrapper filter PSO combinig with SVM.

4. Step of classification:

To test the performance of the genes chosen from the method of selection that have been previously operated, either supervised classification we have used support vector machine learning. (see the chapter 2) and unsupervised classification the clusters produced by the k-means procedure are sometimes called "hard" or "crisp" clusters, since any feature vector x either is or is not a member of a particular cluster. This is in contrast to "soft" or "fuzzy" clusters, in which a feature vector x can have a degree of membership in each cluster. The fuzzy-k-means procedure of Dunn and Bezdek [56] allows each feature vector x to have a degree of membership in Cluster i. Performance of any classifier needs to be tested with some metric, to assess the result and hence the quality of the algorithm. In our study, to evaluate the results of the experiments of machine learning algorithms, we utilized two widely used metrics, i.e. classification accuracy (ACC), area under the Receiver Operating Characteristic Curve (AUC).

Most of the CADx problems deal with two class predictions to map data samples into one of the groups, i.e. benign or malignant. For such a two-class problem, the outcomes are labeled as positive (p) or negative (n). The possible outcomes with respect to this classification scheme is frequently defined in statistical learning as true positive (TP), false positive (FP), true negative (TN) and false negative (FN). These four outcomes are connected to each other with a table that is frequently called

as confusion matrix. For a binary classification scheme, confusion matrix is used to derive most of the well known performance metrics such as sensitivity, specificity, accuracy, positive pre-diction value, and AUC and ROC curve. These metrics are evaluated using the confusion matrix outcomes, i.e. TP, FP, TN and FN predictive values.

Accuracy (Acc): is a widely used metric to determine class discrimination ability of classifiers and it is calculated using equation (1)

$$\textbf{Accuracy (\%)} = \frac{Tp+Tn}{p+n} \qquad (1)$$

It is one of primary metrics in evaluating classifier performances and it is defined as the percentage of test samples that are correctly classified by the algorithm.

Area under the curve (AUC): an important classification performance measure is widely used to measure classifier performances with relevant acceptance. AUC value is calculated from the area under the ROC curve. ROC curves are usually plotted using true positives rate versus false positives rate, as the discrimination threshold of classification algorithm is varied. In terms of TP, FP, TN and FN predictive values, AUC is calculated using Eq. (2).

$$\textbf{AUC} = \frac{1}{2}\left(\frac{Tp}{Tp+Fn} + \frac{Tn}{Tn+Fp}\right) \qquad (2)$$

5. Results:

To evaluate the performance of the proposed method, the selected feature subsets were evaluated by SVM. The result of the experiments for breast dataset is discussed in Table 3.4. The results obtained from base classifier using two metrics like accuracy and ROC.

	With all the attributes 7129 human gene			**Our proposed method 11 human gene**		
Learning classifier Support Vector Machine (SVM)	Acc %	Sensitivity %	ROC	Acc %	Sensitivity %	ROC
50% train_data 50% test_data	62.50	85.71	0.57	93.84	92.18	0.93
70% train_data 30% test_data	66.66	88.88	0.61	93.98	92.43	0.93
80% train_data 20% test_data	68.75	88.89	0.65	98.50	96.77	0.98

Table 3.4: shows the results obtained for the breast dataset using the support vector machine.

We applied K_means clustering so each example belongs to only one cluster (Table 3.5) however in fuzzy K_means clustering, each example is allowed to belong several clusters by using membership values. Membership values mean the degree of distance between centroid and each example (Figure 3.3).

	# Points Classified Incorrectly					
	With all the attributes		**Our proposed method**		**Selected data with our proposed method adding two missing value**	
	K_means	**Fuzzyk_means**	**K_means**	**Fuzzyk_means**	**K_means**	**Fuzzyk_means**
	10	4	19	/	9	2
Correction ratio	0.7959	0.9183	0.6123	/	0.8164	0.9519

Table 3.5: K_means and fuzzy K_means clustering for breast data.

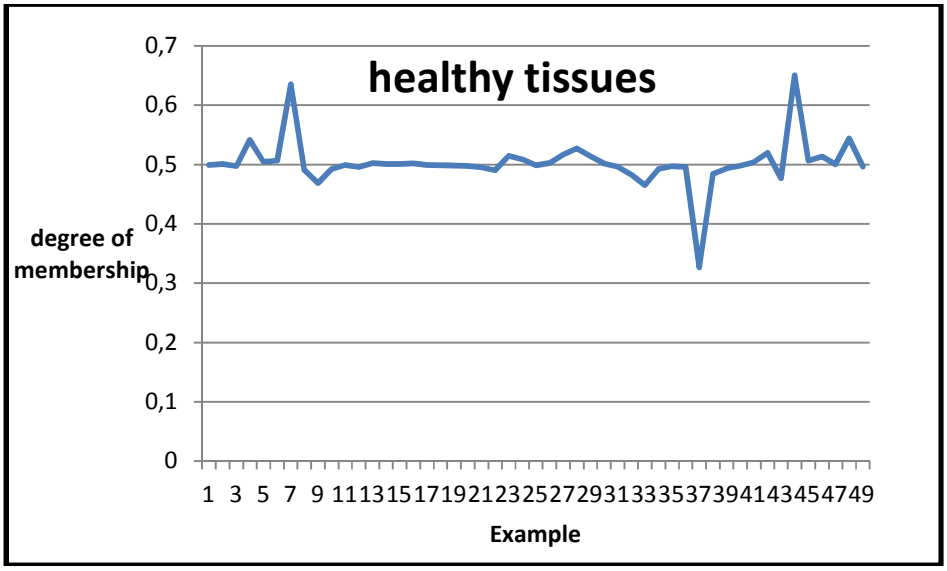

Figure 3.3.a : shows the relationship between the examples and healthy_case.

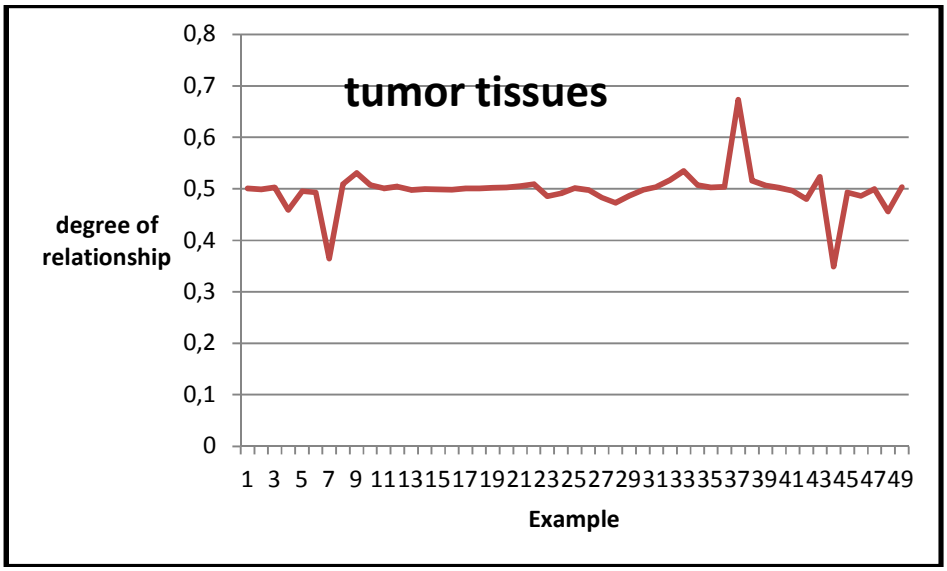

Figure 3.3.b: shows the relationship between the examples and tumor tissues.

Figure 3.3: shows Membership values between the examples and two clusters using fuzzy K_means of all data.

5.1 Discussion and analysis of result:

Several observations can made based on the above experiments, so we tackle the analysis of results focusing on the performance and robustness of our algorithms, as well as the quality of the obtained solutions providing a biological description of most significant ones.

We used PSO to perform feature selection and then evaluated fitness values(extract relevant feature and we guarded hight accuracy) with a SVM, for the SVM configuration the Kernel function was configured as Radial basis function according to our problem of non linear data of breast cancer this function it gives us better result than other variety of function for that our wrapper method PSO_SVM has selected eleven gene belongs to a set of genes (see the biological signification in Table 3.3) with 89.55% accuracy. And about the actors c1 (cognitive learning factor), c2 (social learning factor) is very important. If the parameter adjustment is too small, the particle movement is too small. This scenario will also result in useful data, but is a lot more time-consuming. If the adjustment is excessive, particle movement will also be excessive, causing the algorithm to weaken early, so that a useful feature set cannot be obtained.

selected gene subsets were used to evaluate the performance of classification support vector machine on aspect:

- The number of genes selected by the wrapper filter PSO_Svm and their impact on the classification rate.

In first experimental when we were dividing the base 80% for learning and 20% for testing. The results are shown that the accuracy is 98.50% , this accuracy is best comparing with other (see the Table 3.4) however the high dimensional issue in machine learning has became a big hurdle for the support vector machine classifier itself due to the degradation of classification result that ‘s why the accuracy using all the data is 68.75%.

So the proposed method can serve as an ideal pre-processing tool to help optimize the feature selection process.

In second experimental when we were used the clustering the K-means algorithms are the simplest uses a set of unlabeled feature vectors and classifies them into k classes, where $k = 2$ is given by the user. Algorithm uses distances from the centers of clusters to determine which sample belongs to which class and defined them. The uses of data with all features (7129 genes) give us this correction ratio 0.7959 with ten point classified incorrectly, however the uses of small data (#selecteddata) give

us the correction ratio 0.6123 with nineteen points classified incorrectly.So a comparison of algorithms on real data sets gives a great deal of insight as to their relative performance real data is not worst-case, implying that neither the asymptotic performance or high running time.

We have seen that this result has been approximate to exact case. For that we used Fuzzy Logic because it is suit well to clustering problems so the correction ratio of the data when we combinig the fuzzy logic with K_means is 0.9183 with four point classified incorrectly, the results are determined by some degree of closeness to tumor tissue or to healthy tissue. (See the Figure 3.2) so Fuzzy Logic has been widely used to provide flexibility. Moreover the application of fuzzy logic combinig with K_means to our selected data is failed. It determine that even example have the same degree to close on tumor tissue or healthy tissue with degree of membership 0.50 for two reasons:

- The first, the result is far than the exact case (based on theory research).
- The second, the best way to evaluate the performance of our selected method is the contribution with the biologist, i summarized during my contribution that my selected data have two missing value and fuzzy logic have used to treat the missing value for that it give us the same degree to close on tumor tissue or healthy tissue. Meanwhile, the data set with missing values is another challenge to be conquered.

For that the ad of missing value (BRCA1 and BRCA2) gives us better result comparing with other (See the Table 3.5).

6. Summary of the selection techniques:

We summarize in Table 3.6 the different characteristics of each method during its operation and its treatment during the selection and classification process.

The methods	**Advantages**	**Inconvenients**
Wrapper filter Particle swarm optimization combinig with support vector machine (Pso_Svm)	-The PSO-SVM takes the advantage of minimum structural risk of SVM and the quick global optimizing ability of PSO.	-The application of the algorithm of optimization is influenced by factors such as the criterion of stop, the structure of particle. -More costly in computation time (repetitive executions learning algorithms on different subset of variable)
Support vector machine (SVM)	-It has a regularisation parameter. - It uses the kernel trick, so you can build in expert knowledge. - SVM is defined by a convex optimisation problem (no local minima).	- The determination of the parameters for a given value of the regularisation and kernel parameters and choice of kernel.
K_means	- K_means is one of the simplest algorithm which uses unsupervised learning method to solve know clustering issues. It works really well with large datasets. - Esaier to understand	- Number of cluster and initial seed value need to be specified beforehand -Low capability to pass the local optimum.
Fuzzy K_means	-Gives best result for overlapped data set and comparatively better than k-means algorithm. -Here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.	-Apriori specification of the number of clusters.

Table 3.6: characteristics of each method of the selection and classification process.

7. Conclusion:

Our variable selection method was used to build predictors effective for a bipartition problem supervised or unsupervised expression data DNA chip (microarray) in cancer (breast cancer). The performance obtained are as good because The wrapper model differs from the filter model in that it is dependent on a classifier and evaluates the combination of feature subsets.The wrapper model can identify interaction amongst all features simultaneously. This characteristic is important for the robustness of our predictors with a necessary condition for a possible use in clinical routine. And fuzzy logic kmeans formularize an intuitive theory in problems where the results can be approximate rather than exact.

So it could also be applied to problems in other areas in the future in order to build an efficient classification model for classification problems with different dimensionality and different sample size is important.

Conclusion and feature works:

Our aim in this work was to develop new tools for breast cancer management to help the physicians in their decision-making practices. In this order an attempt to propose suitable approaches has been performed within machine learning framework, to enable handling the main recent challenges encountered in breast cancer management field. Some challenges are due to the intrinsic complexity of data issued from high throughput technologies introduced recently in cancer management such as microarrays. The gene expression profiling, through microarray technology, has indeed brought the hope to gain new insights into cancer biology but requires mean while smart approaches capable to fit with high dimensional data and uncertainties. Uncertainties can be in the form of either measurement noise or membership. Another challenge is related to the use of traditional clinical factors characterized by its heterogeneity, the data can be of quantitative or symbolic type.

In a first work a wrapper feature selection approach based on support vector machine learning able to deal with high dimensional data has been proposed. This approach proposes a new algorithm to solve this problem in the primal domain. The basic idea is to optimize the learning classifier using partical swarm optimization. It has been shown through large-scale numerical experiments that the proposed approach is computationally more efficient than the few existing methods solving the same problem.

However, with the recent trends towards an integrative bioinformatics that aims to integrate different data sources, the occurrence of three challenges simultaneously is possible in some cancer applications. To deal simultaneously with these three challenges; data dimensionality, heterogeneity and uncertainties.

A unified principle to deal with data heterogeneity problem has been established. To take into account membership uncertainty and increase model interpretability, this principle has been proposed within a unsupervised fuzzy logic framework in order to develop a new weighted fuzzy rule-based clustering algorithm. An extensive study has been also performed to compare this algorithm with one of the state-of-the-art clustering algorithm.

This area of research will remain active as long and motivated by changes systems for collecting and storing data on the one hand and the requirements on the other hand. The best approach for judging this selection is collaborating with experts (biologists) for an interpretation of the results and highlight the following:

- Genes that are involved in susceptibility and of these cancers.
- The genes that contribute to tumor development of breast cancers.

This collaboration with biologists allows us how to use these fundamentals in clinical practice and their influence on decision of patients because it is a major area of research in: screening, treatment and predicting the clinical course of these patients.

In parallel, we aime to better understand the evolution of the patient's immune response for all stages of development of breast cancer.

Appendix A:

Most cases of breast cancer are not inherited. These cancers are associated with genetic changes that occur only in breast cancer cells (somatic mutations) and occur during a person's lifetime.

In hereditary breast cancer, the way that cancer risk is inherited depends on the gene involved. For example, mutations in the BRCA1 and BRCA2 genes are inherited in an autosomal dominant pattern, which means one copy of the altered gene in each cell is sufficient to increase a person's chance of developing cancer. (See Table 1).

Protein /Gene	**Abbreviation**	**Description**
Breast cancer 1, early onset.	BRCA1	The BRCA1 gene belongs to a class of genes known as tumor suppressor genes. The BRCA1 gene provides instructions for making a protein that is directly involved in repairing damaged DNA; BRCA1 plays a role in maintaining the stability of a cell's genetic information. the BRCA1 protein also regulates the activity of other genes and plays a critical role in embryonic development.
Breast cancer 2, early onset.	BRCA2	The BRCA2 gene belongs to a class of genes known as tumor suppressor genes. the protein produced from the BRCA2 gene helps prevent cells from growing and dividing too rapidly or in an uncontrolled way. The BRCA2 gene provides instructions for making a protein that is directly involved in the repair of damaged DNA.
ATM serine/threonine kinase	ATM	The ATM protein coordinates DNA repair by activating enzymes that fix the broken strands. Efficient repair of damaged DNA strands helps maintain the stability of the cell's genetic information.
tumor protein p53	P53	This protein acts as a tumor suppressor, which means that it regulates cell division by keeping cells from growing and dividing too fast or in an uncontrolled way. P53 is essential for regulating cell division and preventing tumor formation.

checkpoint kinase 2	CHEK2	This protein acts as a tumor suppressor, which means that it regulates cell division by keeping cells from growing and dividing too rapidly or in an uncontrolled way. This process keeps cells with mutated or damaged DNA from dividing, which helps prevent the development of tumors.
cadherin 1, type 1, E-cadherin (epithelial)	CDH1	Whose function is to help neighboring cells stick to one another (cell adhesion) to form organized tissues. which means it prevents cells from growing and dividing too rapidly or in an uncontrolled way.
serine/threonine kinase 11	STK11	This protein aids in the prevention of tumors, especially in the gastrointestinal tract, pancreas, cervix, ovaries, and breasts. This protein function is also required for normal development before birth
partner and localizer of BRCA2	PALB2	This protein interacts with the protein produced from the BRCA2 gene. These two proteins work together to mend broken strands of DNA, which prevents cells from accumulating genetic damage that can trigger them to divide uncontrollably the PALB2 and BRCA2 proteins play a critical role in maintaining the stability of a cell's genetic information.

Table 1: genes are related to breast cancer [53].

Appendix B:

In this section, we describe the proposed SVM-PSO classification system for the classification of high-dimensional data. As mentioned in the Introduction, the aim of this system is to optimize the SVM classifier accuracy by automatically.

The process of PSO algorithm for solving optimization problems

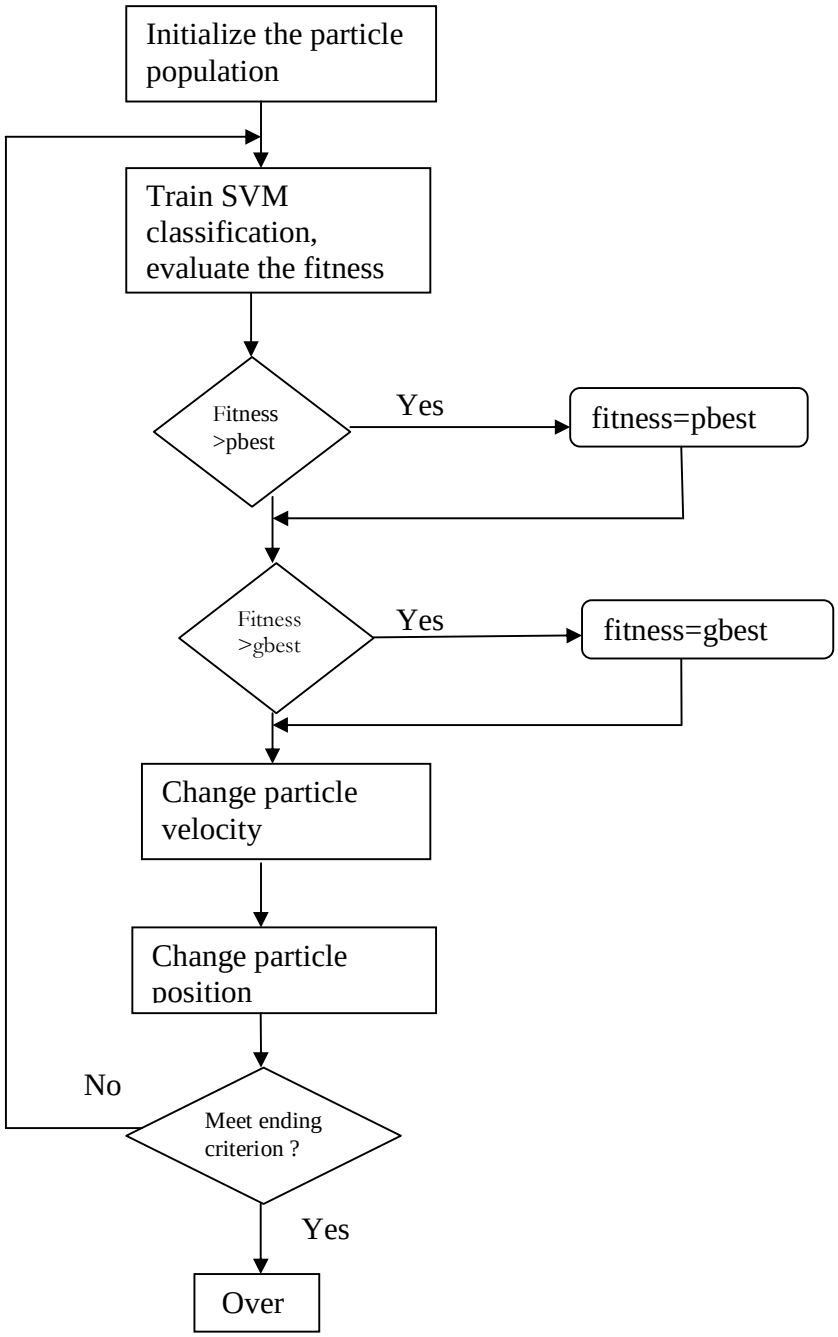

Figure1: The process of PSO algorithm for solving optimization problems.

References:

[1] Impact of breast cancer, http://www.who.int . (Consult 02-02-2015)
[2] The normal structure of the breasts, http://www.cancer.org/cancer/breastcancer. (Consult 04-02-2015)
[3] http://www.breastcancer.org/symptoms/understand_bc/what_is_bc. (Consult 04-02-2015)
[4] http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-treating-by-stage . (Consult 04-02-2015)
[5] http://www.cancer.ca/en/cancer-information/cancer-type/breast/prognosis-and-survival/survival-statistics/region=on. (Consult 04-02-2015)
[6] http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics . (Consult 05-02-2015)
[7] http://www.quora.com. (Consult 12-02-2015)
[8] http://openi.nlm.nih.gov. (Consult 12-02-2015)
[9] Cruz J.A., Wishart D.S., Application of Machine Learning in Cancer Prediction and Prognosis, Cancer Informatics, 2, pp. 59-77, 2006.

[10] Miller L.D., Smeds J., George J., An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival, PNAS September 20, 102 (38), pp. 13550-13555, 2005.

[11] Michiels S., Koscielny D., Hill C., Prediction of cancer outcome with microarrays: a multiple random validation strategy, The Lancet, 365 (9458), pp. 488-492, 2005.

[12] Reid J.F. , Lusa L., De Cecco L., *and al.*, Limits of Predictive Models Using Microarray Data for Breast Cancer Clinical Treatment Outcome, JNCI J Natl Cancer Inst , 97 (12), pp. 927-930, 2005.

[13] Sun Y., Goodison S., Li J., Liu L., Farmerie W., Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics, Gene expression, 23 (1), 30-37, 2007a.

[14] Fukunaga K., *Introduction to statistical pattern recognition*, 2nd ed. New York: Academic, 1990.

[15] Parry R.M., Jones W., Stokes T.H., *et al.*, *k*-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction, Bioinformatics, 18(7), pp. 961-970,2002.

[16] Olshen A.B., Jain A.N., Deriving quantitative conclusions from microarray expression data, Bioinformatics,18 (7), pp. 961-970, 2002.

[17] Zheng B., Wang X., Lederman D., Tan J., Gur D., Computer-aided detection; the effect of training databases on detection of subtle breast masses, Pharmacogenomics J., 10(4), pp. 292-309, 2010.

[18] Baldi P., Brunak S., Bioinformatics: The machine learning approach, MIT Press, Cambridge,2001.

[19] https://randomforests.wordpress.com. (Consult 12-02-2015)
[20] Liu H.X., Zhang R.S., Luan F., *et al.*, Diagnosing Breast Cancer Based on Support Vector Machines, J. Chem. Inf. Comput. Sci., 43 (3), pp 900–907, 2003.

[21] Chang R-F., Wu W-J., Moon W.K., *et al.*, Support Vector Machines for Diagnosis of Breast Tumors on US Images, Academic Radiology, 10 (2), pp. 189-197, 2003b.

[22] Land W.H., Verheggen E.A., Multiclass primal support vector machines for breast density classification, Int J Comput Biol Drug Des., 2(1), 21-57, 2009.

[23] Baldi P., Brunak S., *Bioinformatics: The machine learning approach*, MIT Press, Cambridge,2001.

[23] www.cs.umd.edu . (Consult 12-02-2015)
[24] http://blog.peltarion.com. (Consult 16-02-2015)
[25] Filippone M., Camastra F., Masulli F., Rovetta S., A survey of kernel and spectral methods for clustering, Pattern recognition, 41, pp. 176-190, 2008.

[26] Wang J., Bø T.H., Jonassen I., *et al.*, Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data, BMC bioinformatics, 4(60), 2003.

[27] Wiseman S.M., Makretsov N., Nielsen T.O., *et al.*, Coexpression of the Type 1 Growth Factor Receptor Family Members HER-1, HER-2, and HER-3 has a Synergistic Negative Prognostic Effect on Breast, Cancer, 23 (9), pp. 1770-1777, 2005.

[28] Covell D.G., Wallqvist A., Rabow A.A., Thanki N., Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data, Mol Cancer Ther, *2*, pp. 317-332, 2003.

[29] De Souto M., Costa I., De Araujo D., *et al.*, Clustering cancer gene expression data: a comparative study. BMC Bioinformatics, 9(1), 497, 2008.

[30] George H . John, Ron Kohavi, and Karl Pfleger. Feature and the subset selection problem. In international conference on Machine learning. Journal version in 1IJ, pages 121_129, 1994
[31] P.Langley .Selection of relevant features in machine learning .In Proceeding of the AAAI Fall symposium on Relevance A AAI press,1994.
[32] G. H. John, R. Kohavi, K. Pfleger, “Irrelevant feature and the subset selection problem,” in Proc. of the Eleventh International Conference on Machine Learning, pp. 121-129, 1994.

[33] Guf .S. Unsupervised feature selection: when random ranking sound are irrelevancy. In JMCR workshop and conference proceding, New challenges for feature selection in data mining and knowledge discovery, 4:161_175, 2008.

[34] C.A Murthy Mitra and S.K. Pal. Unsupervised feature selection using
feature similarity. IEEE Trans. On patterns Analysis and machine
learning, pages 24_34, 2002.
[35] M. Hall. Correlation-based feature selection for machine learning. 1998.
[36] J. Doak, “An evaluation of feature selection methods and their application to computer security,” Technical report, Davis CA: University of California, Department of Computer Science, 1992.
[37] P. Narendra, K. Fukunaga, “A Branch and Bound Algorithm for Feature Subset Selection,” IEEE Transactions on Computer, vol. 26, no. 9, pp. 917-922, 1977. Article (CrossRef Link) ,” J. Pearl, “Heuristics,” Addison-Wesley, 1983.

[38] H. Liu, H. Motoda, “Feature Selection for Knowledge Discovery and Data Mining,” Kluwer Academic Publishers, London, GB, 1998.

[39] Huan Liu, Lei Yu, “Toward Integrating Feature Selection Algorithms for Classification and Clustering,” IEEE Transaction on Knowledge and Data Engineering, 2005.

[40] M. Dash, H. Liu, “Feature Selection for Classification,” Intelligent Data Analysis, Elsevier, pp. 131-156, 1997.

[41] Yuhang Wing and Fillia Makedon. Application of relie_ feature _ltering algorithm to selecting informative genes for cancer classi_cation using microarray data. IEEE Computational Systems, pages 497_498, 2004.
[42] Tian Lan, Deniz Erdogmus, Andre Adami, and Michael Pavel. Feature selection by independent component analysis and mutual information maximization in eeg signal classi_cation. An International Journal,170 :409_418, 2005.

[43] Shoushan Li, Rui Xia, Chengqing Zong, and Chu-Ren Huang. A framework of feature selection methods for text categorization. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2-7 :692700, 2009.

[44] Mehmet Fatih Akay *.2009Support vector machines combined with feature selection for breast cancer diagnosis . Expert Systems with Applications 36 (2009) 3240–3247 .

[45] Prasad, Y., Biswas, K., & Jain, C. (2010). Svm classifier based feature selection using ga, aco and pso for sirna design. In Proceedings of the first international conference on advances in swarm intelligence (pp. 307–314).

[46] Mohammad Javad Abdi,SeyedMohammd Hosseini, and Mansoor Rezghi., A NovelWeighted Support Vector Machine Based on Particle Swarm Optimization for Gene Selection and Tumor Classification, Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2012, Article ID 320698, 7 pages doi:10.1155/2012/320698

[47] Bichen Zheng, Sang Won Yoon , Sarah S. Lam.2014 Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications 41 (2014) 1476–1482.

[49] W Du; Y Sun; Y Wang; etal.,Int. J. Data Mining and Bioinformatics,2013, 7(1), 58-77.
[50] Sun Y., Goodison S., Li J., Liu L., Farmerie W., Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics, Gene expression, 23 (1), 30-37, 2007a.

[51] Gevaert O., De Smet F., Timmerman D., Moreau Y., De Moor B., Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian network, Bioinformatics, 22 (14), 184-190, 2006.

[52] West M, & al.2001 Predicting the clinical status of human breast cancer using gene expression profilesProc. Natl Acad. Sci. USA 98 11462 11467. doi:10.1073/pnas.201162998 .

[53] http://ghr.nlm.nih.gov/condition/breast-cancer. (Consult 22-02-2015)

[54] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," 2005. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[55] gene bank. http ://www.ncbi.nlm.nih.gov. (Consult 22-04-2015)

[56] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York, 1981). A clearly written monograph that emphasizes fuzzy clustering.