VALIDATION STUDY OF DIMENSIONALITY REDUCTION IMPACT ON BREAST CANCER CLASSIFICATION

Nezha Hamdi*, Khalid Auhmani**, Moha M'rabet Hassani*

*Department of Physics, Faculty of sciences Semlalia, Cadi Ayyad University
Marrakech, Morocco

** Department of Industrial Engineering, National school of applied sciences
Cadi Ayyad University, Safi, Morocco

ABSTRACT

A fundamental problem in machine learning is identifying the most representative subset of features from which we can construct a predictive model for a classification task. This paper aims to present a validation study of dimensionality reduction effect on the classification accuracy of mammographic images. The studied dimensionality reduction methods were: locality-preserving projection (LPP), locally linear embedding (LLE), Isometric Mapping (ISOMAP) and spectral regression (SR). We have achieved high rates of classifications. In some combinations the classification rate was 100%. But in most of the cases the classification rate is about 95%. It was also found that the classification rate increases with the size of the reduced space and the optimal value of space dimension is 60. We proceeded to validate the obtained results by measuring some validation indices such as: Xie-Beni index, Dun index and Alternative Dun index. The measurement of these indices confirms that the optimal value of reduced space dimension is d=60.

KEYWORDS

Dimensionality reduction, Classification, Validation indices, K-nearest neighbors, Machine learning.

1. INTRODUCTION

Dimensionality reduction problem has been studied by scientific communities for statistics and machine learning purposes for many years [1, 2]. It has received more attention recently because of the promising results in data mining. It is not also surprising that dimensionality reduction is an enthusiastic research area for machine learning and for pattern recognition; both fields share the common task of classification.

The dimensionality reduction problem often arises when it comes to consider a very large number of variables. In recent years the need has evolved with the manipulation of very large databases and especially in areas such as genetic field and image processing field [3]. Consequently the number of features should be reduced.

Dimensionality reduction methods try to find a projection of the data in a new space of reduced dimension, without losing the information. This projection may be linear or nonlinear. They are generally classified into three categories: the Wrapper, Filter, and Embedded [4]. The Wrapper approaches use the classification error rate as a evaluation criteria [5]. They then incorporate the classification algorithm in the search and selection of attributes. These methods allow the obtaining of high performance. However, the use of such methods requires for each subspace of

attributes to perform classification, which can become costly in calculation time especially when the dimension d of the input space is large. These methods are very dependent of the used classification algorithm.

Filter approaches use an evaluation function based on the characteristics of all data, independently of any classification algorithm [6-10]. These methods are fast, general and less expensive in computation time, which allows them to operate more easily with databases of very large dimensions. However, as they are independent of the classification stage, they do not guarantee to reach the best classification accuracy.

In order to combine the advantages of both methods, hybrid algorithms "embedded" have been proposed. The dimensionality reduction process is performed in conjunction with the classification process. A filter-type evaluation function is first used to screen the most discriminating feature subspace. Then the error rates of misclassification, by considering each discriminant subspace previously selected, are compared in order to determine the final subspace [11, 12].

Due to their computational efficiency and independence of any classification algorithm, the “filter” approaches are more popular and commonly used. The application of cluster analysis has been demonstrated to be more effective than traditional dimensionality reduction algorithms [3].
Our goal in this paper is to perform a validation study of the dimensionality reduction methods effect on the classification accuracies of breast cancer (mammographic image). The quality of such dimensionality reduction process determines the purity of classification and hence it is very important to evaluate the results of the dimensionality reduction algorithms. Due to this, validation of reduced space dimension had been a major and challenging task. The main goal of this paper is to measure some validation indices such as Xi-Beni, Dunn and Alternative Dunn indices

2. MATERIALS AND METHODS

2.1. Features extraction.

Initially, features are calculated to form the feature vector for subsequent learning step. Features were extracted from a set of tow classes labeled images). Images are firstly pre-processed and transformed by the discrete double density wavelet transform (3DWT). The following features are extracted: Texture descriptors [13], - Statistical moments [14], Tamura parameters [15,16], Radon’s characteristics [17, 18] and Zernike’s moments[19]

2.2.Features selection

For dimensionality reduction task we will test the following methods:

Locality Preserving Projections (LPP): A graph incorporating neighborhood information of the data set is built [20]. This linear transformation optimally preserves local neighborhood information in a certain sense. The generated map may be viewed as a linear discrete approximation to a continuous map [21].

Locally linear embedding (LLE) applies dimensionality-reduction to the data for learning and classification. The objective of this method is to determine a locally-linear fit, so that each data point can be represented by a linear combination of its closest neighbors [22]. It consists of three main steps:

- Find the K nearest neighbors of each D-dimensional input data point Xi , i = 1, ..., N. The Euclidean distance is used as a similarity measure.
- Calculate the weights Wij that best reconstruct each data point Xi from its neighbors by minimizing the following equation:

$$\varepsilon(W) = \sum_{i=1}^{N} \left\| X_i - \sum_{j=1}^{K} W_{ij} X_j \right\|^2$$

- Calculate the low-dimensional embedding Yi . The weights Wij are kept fixed and the following cost function is minimized:

$$\phi(Y) = \sum_{i=1}^{N} \left\| Y_i - \sum_{j=1}^{K} W_{ij} Y_j \right\|^2$$

Isometric Feature Mapping has been introduced by Tenenbaum [23]. This is a non-linear extension of multidimensional scaling (MDS) [24]. The procedure also consists of three main steps:

- Search the K nearest neighbors for each data point Xi.
- Build the neighborhood graph G and calculate the shortest path dG(i; j) between any two given data points. Construct an embedding of the data by applying classical MDS to the matrix of graph distances DG ={dG(i;j)}.

Spectral regression (SR): the background of this method can be consulted in [25,26].

2.3.Benchmark

Our benchmark (Figure 1) is divided in four parts. The original image will first be preprocessed to reduce noise and enhance the presentation. In the next part we apply the discrete double density wavelet transform to the image. The third step deals with the extraction of features as described above. In the fourth part, the studied Dimensionality reduction methods will be applied to the complete dataset. The low-dimensional datasets will then be classified by the K Nearest Neighbors classifier. We will compare the performance of each method. The classification accuracies of KNN will be plotted versus reduced dimension. Finally we will measure the validation indices. The studied methods were tested on the MIAS images [27].

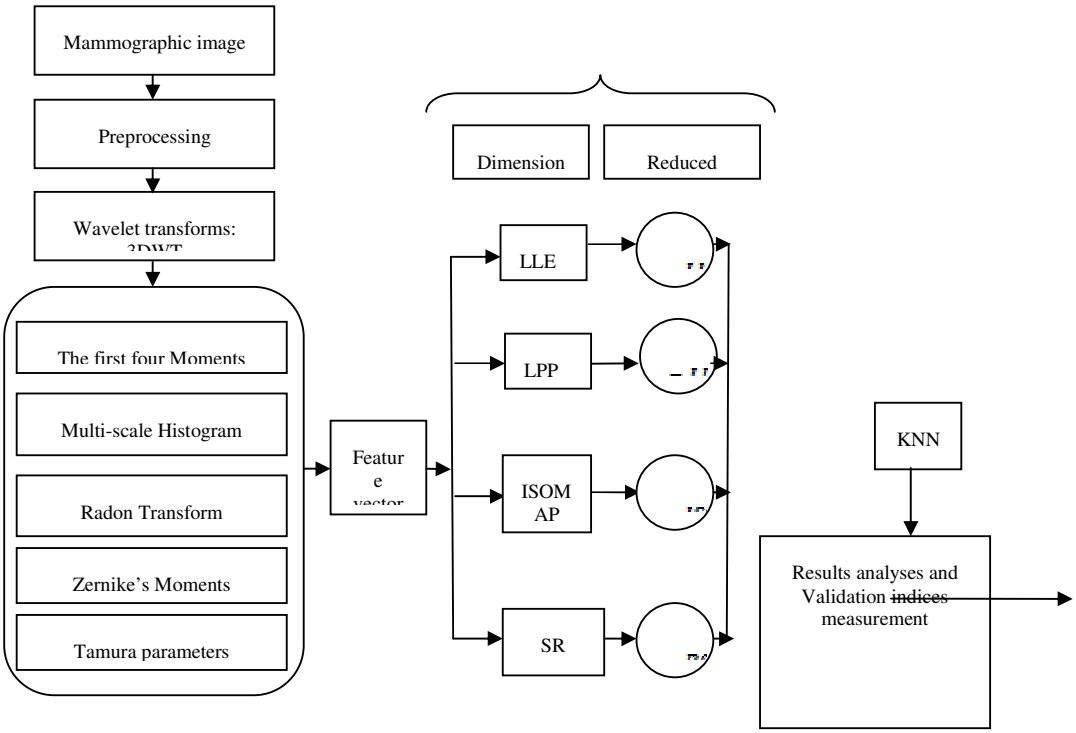

Figure 1 Benchmark of the propposed system

3. RESULTS AND DISCUSSION

3.1. Dimension influence

Figure 2.presents the classification accuracy versus reduced space dimension (d). Features are extracted from transformed image by the double-density wavelet transform (3DWT). We can see that the performance has reached 100% for Spectral regression method (SR) and for space dimension d=60. However, the comparison of the accuracies corresponding to LLP, LLE and ISOMAP, shows that reducing the space for small size (d=5) the classification accuracy of 95.3% is held

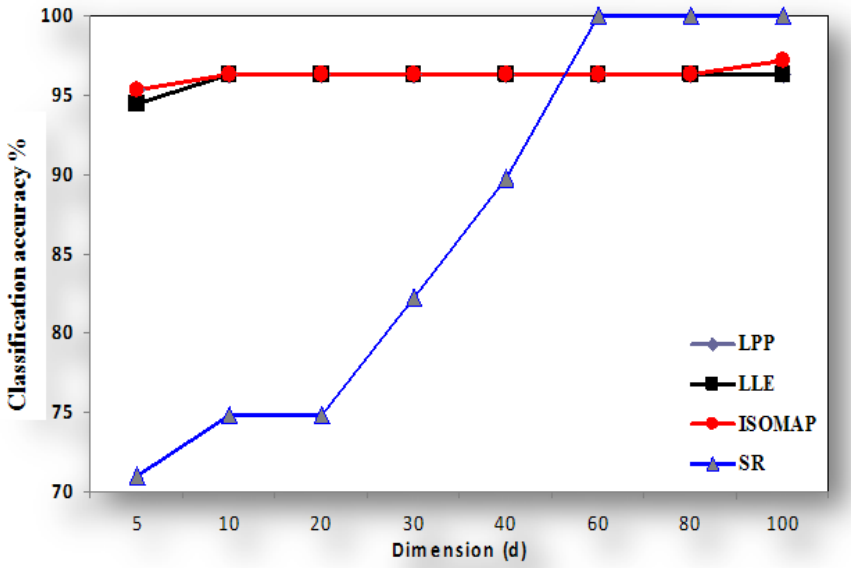

Figure 2 Effect of different Dimensionality reduction methods on classification accuracy, image ransformed by 3DWT.

3.2.Validation indices measurement

After identifying the partitions by the previous step, the question immediately arises: what is the optimum distance to choose for a good partition. The answer to this question is included in the field of cluster Validation assessment. The goal of clusters validation techniques is to measure the best partition relative to others obtained by other partitioning algorithms or by using the same algorithms but with different parameters. The evaluation and the validation of the optimal discrimination is based on two criteria [28, 29]:

Compactness: It measures the uniformity and consistency of data in each class. Data from the same class generate highly compact partition. The evaluation of the compactness of a partition depends on the used measure. For example, if the variance is measured, a minimum value means a great compactness. Conversely, if the average similarity is used; more this value is higher more the partition is compact [30].

Separation: A "good" partitioning means classes are well separated. Measuring the separation between the two classes can be performed in three ways: (i) - measuring the distance between the closest data of the two classes, (ii) - by measuring the distance from the most distant data and (iii) - measuring the distance between cluster centers. A large value of this distance leads to good separation. If the similarity between classes is used, more this value is weak, more the classes are well separated [31].

For the presentation of the different indices, the following notations are adopted: c is the number of clusters, n the number of features, χ_i the iith cluster, v_i the center of iith cluster and $d(x, y)$ the distance between two objects.

Dunn index (DI):

One way to assess the quality of a group is to compare its dispersion to the distance between the nearest groups. Indeed, if the inter-distance is larger than the dispersion of objects within a group, these groups are disjoint. The ratio of these distances is a good indicator of the group quality.

Dunn index [32], takes this approach by calculating the ratio of the minimum distance intra-class (diameter classes) and the maximum interclass distance (dissimilarity between classes). Let S and T are two non-empty subsets of $\Re^2$. The diameter Δ of S and distance δ are then:

$$\Delta(S) = \underbrace{\max}_{x, y \in S} \{d(x, y)\}$$

$$\delta(S, T) = \underbrace{\min}_{x \in S, y \in T} \{d(x, y)\}$$

Dunn index is then:

$$v_D = \underbrace{\min}_{1 \le i \le c} \left\{ \underbrace{\min}_{1 \le j \le c, j \ne i} \left\{ \frac{\delta(\chi_i, \chi_j)}{\underbrace{\max}_{1 \le k \le c} (\Delta(\chi_k))} \right\} \right\}$$

The main objective of this index is to maximize the inter-cluster distance and minimize the intra-cluster distance. The objective is to maximize the index Dunn.

Alternative Dunn index (ADI)

The calculation of the original Dunn index becomes easier when the dissimilarity between cluster centers respects this relationship [33]:

$$d(x, y) \ge \left| d(y, v_j) - d(x, v_j) \right|$$

As v_j is the center of cluster j, the alternative Dunn index is:

$$ADI(c) = \min_{i \in c} \left\{ \min_{j \in c, i \ne j} \left\{ \frac{\min_{x_i \in C_i, x_j \in C_j} \left| d(y, v_j) - d(x_i, v_j) \right|}{\max_{k \in c} \left\{ \max_{x, y \in C} d(x, y) \right\}} \right\} \right\}$$

Xie-Beni index: V_{XB}

It provides a Validation measure of compactness and separability of groups. A small value of V_{XB} indicates an optimal partitioning [34]:

$$V_{XB} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^2 d^2(x_i, v_j)}{n [\min_{k \ne j} (d^2(v_k, v_j))]}$$

The numerator represents the compactness measured by the sum of the squared distances intra classes while the denominator represents the measured separation by the minimum interclass's distance. The Xie-Beni index will be a small value when the partition is good.

Evaluation and validation of different partitions are based on the Validation indices presented above. For each of these indices, we accept that dimension is optimal for good classification. All the results are presented in Figures 3 and 4. The results are in coherence with the quantitve results:

For the V_{XB} index that measures the compactness of the cluster, we find that the best value (minimum) for this test was obtained for d = 60 with the Spectral regression method.
For Dunn index that measures the distance between the obtained clusters, we notice that it finds its best value (maximum) for d = 60.

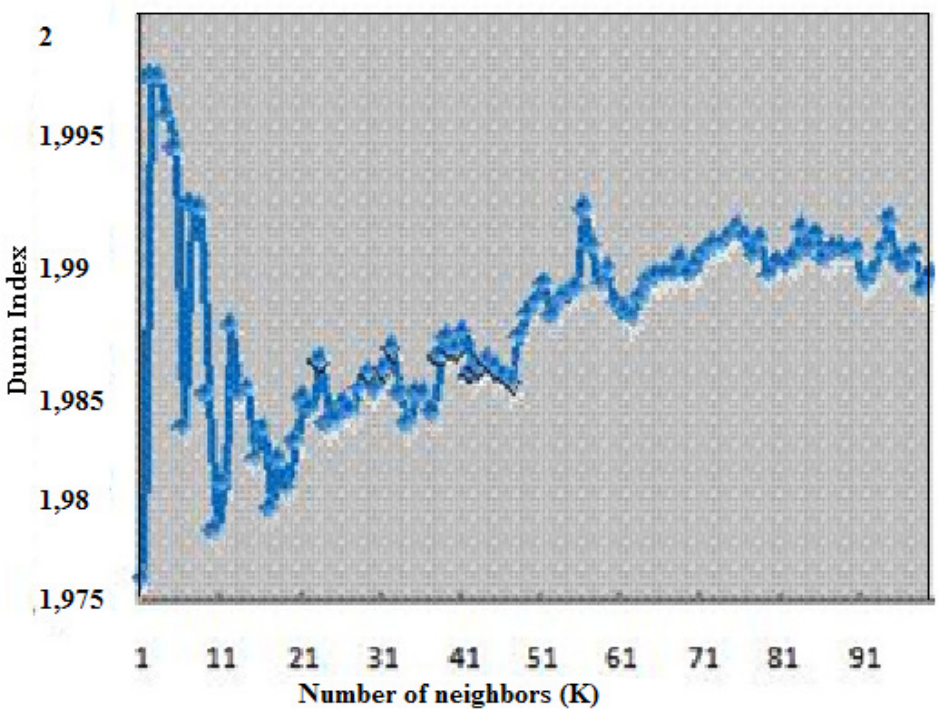

Figure 3 Dunn index versus the k nearest neighbours

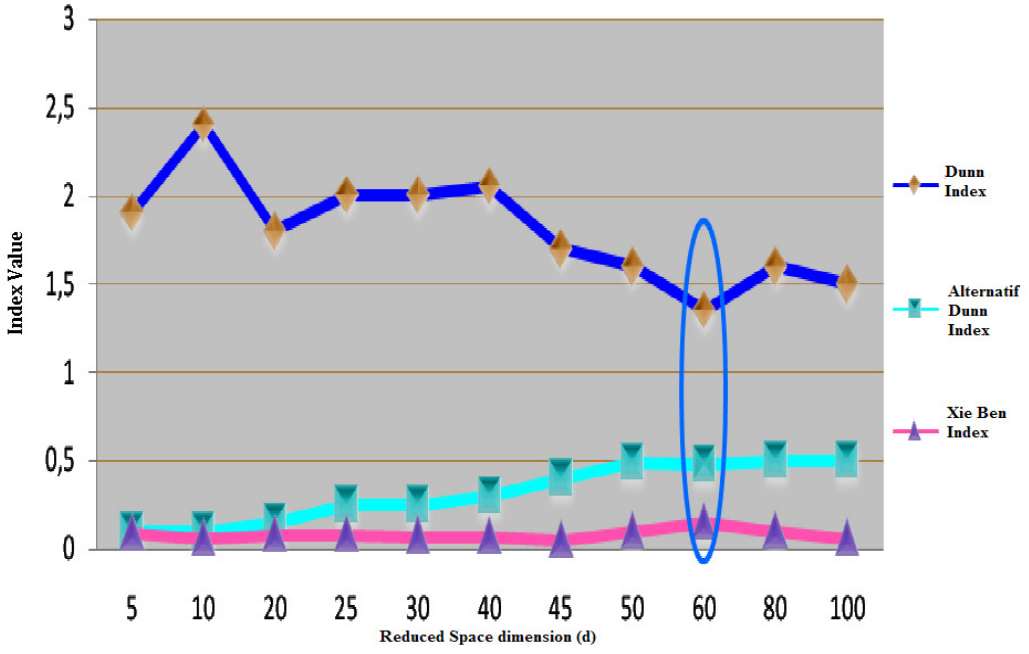

Figure 4 Validation indices versus reduced space dimension

4. CONCLUSION

The dimensionality reduction problem is justified when the data to be treated are very large. This topic has been widely studied and we have made some contribution concerning classification of mammographic images. Our goal in this paper was to validate our results previously found and reported in our previous works [35, 36].

During the experimental phase, we compared the different techniques of feature selection methods associated with a wavelet transform applied to the image before the features extraction process. We have reached a classification accuracy of 100% for spectral regression method. We also found that generally the classification accuracy increases with the dimension but stabilizes after a certain value which is d=60.

We proceeded to validate the obtained results by measuring certain Validation indices: Xie-Beni Index, Dunn and alternative Dunn indices. The measurement of these indices confirms the quantitative obtained results; the growth of classification accuracy with the dimension and the optimal value of d in the studied case is d = 60.

REFERENCES

[1] P. A. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach. Prentice/Hall International, 1982.
[2] J. Miller. Subset Selection in Regression. Chapman and Hall, New York, 1990.
[3] Ali El Akadi, « Contribution à la sélection de variables pertinentes en classification supervisée: application à la sélection des gènes pour les puces à ADN et des caractéristiques faciales » ; these de doctorat, Faculté des sciences Rabat, 2012.
[4] Qinbao Song, Jingjie Ni and Guangtao Wang; A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data; Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 1), 2011

[5] R.Kohavi and G. John. Wrappers for feature subset selection. Artificial Intelligence, pages 273–324, 1997.
[6] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
[7] Langley P., Selection of relevant features in machine learning, In Proceedings of the AAAI Fall Symposium on Relevance, pp 1-5, 1994.
[8] Souza J., Feature selection with a general hybrid algorithm, Ph.D, University of Ottawa, Ottawa, Ontario, Canada, 2004.
[9] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In Proceedings of the Advances in Neural Information Processing Systems 'NIPS 05', pages 507–514, Vancouver, Canada, December 2005.
[10] L. Talavera. Feature selection as a preprocessing step for hierarchical clustering. In Proceedings of the 16th International Conference on Machine Learning 'ICML 99', pages 433–443, Bled, Slovenia, 1999.
[11] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In Proceedings of the 18th International Conference on Machine Learning 'ICML 01', pages 74–81, Williamstown, MA, USA, June 2001.
[12] Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003.
[13] Hadjidemitriou, E., Grossberg, M. D., & Nayar, S. K.. Multiresolution histograms an their use for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26 (7), 831-847. (2004)
[14] S. HERLIDOU, Caractérisation tissulaire en IRM par l'analyse de texture. Étude du tissu musculaire et de tumeurs intracrâniennes, université de Renne1, 1999.
[15] H. TAMURA, S. MORI et T. YAMAWAKI. Texture features corresponding to visual perception IEEE Transactions on Systems, Man and Cybernetics, SMC-8(6):460–473, 1978.
[16] P. HOWARTH et S. RÜGER. Evaluation of texture features for content-based image retrieval. In Proceedings of the International Conference on Image and Video Retrieval (CIVR'04), volume LNCS 3115, pages 326–334, Dublin, Ireland, jul 2004.
[17] Deans, S.R. Hough Transform from the Radon Transform_, IEEE Trans. On Patt. Anal. and Mach. Intell., Vol. PAMI-3, No. 2, pp. 185_188, 1981.
[18] Murphy, L.M. Linear feature detection and enhancement in noisy images via the Radon transform, Pattern Recognition Letters, No. 4, pp. 279_284, 1986.
[19] C.-W. Chong, P. Raveendran, and R. Mukun-dan. A comparative analysis of algorithms for fast computation of Zernike moment. Pattern Recognition, 36:731-742, 2003.
[20] Xiaofei He, and Partha Niyogi, "Locality Preserving Projections" , Advances in Neural Information Processing Systems 16 (NIPS 2003), Vancouver, Canada, 2003.
[21] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Advances in Neural Information Processing Systems 14, Vancouver, British Columbia, Canada, 2002.
[22] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323-2326 (2000).
[23] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimen-sionality reduction. In Science, volume 290, pages 2319{2323, 2000.
[24] T. Cox and M. Cox. Multidimensional Scaling. Chapman and Hall, London, 1994.
[25] Deng Cai, Xiaofei He, and Jiawei Han. "Semi-Supervised Regression using Spectral Techniques", Department of Computer Science Technical Report No. 2749, University of Illinois at Urbana-Champaign (UIUCDCS-R-2006-2749), July 2006.
[26] Deng Cai, Xiaofei He, and Jiawei Han. "Spectral Regression for Dimensionality Reduction", Department of Computer Science Technical Report No. 2856, University of Illinois at Urbana-Champaign (UIUCDCS-R-2007-2856), May 2007
[27] http://peipa.essex.ac.uk/info/mias.html.
[28] M. J. BERRY et G. LINOFF. Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Inc., New York, NY, USA, 1997.
[29] R. DUBES et A. K. JAIN. Validity studies in clustering methodologies. Pattern Recognition, 11:235–254, 1979.
[30] M. PARTIO, B. CRAMARIUC,M. GABBOUJ et A. VISA. Rock texture retrieval using gray level cooccurrence matrix. In 5th Nordic Signal Processing Symposium (NORSIG'02), Trollfjord, Norway, octobre4-7 2002.

[31] N. IDRISSI, La navigation dans les bases d'images: prise en compte des attributs de texture, thése de doctorat, l'université de Nantes ,2004,
[32] J. C. DUNN. Well separated clusters and optimal fuzzy partitions. Journal of Cybernetica, 4:95–104, 1974.
[33] Balazs Balasko, Janos Abonyi ,Balazs Feil, Fuzzy Clustering and Data Analysis Toolbox, University of Veszprem ,Hungary, 2005
[34] X. L. XIE et G. BENI. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal Mach. Intell., 13(8):841–847, 1991.
[35] N. Hamdi , S. H. Bouazza, K. Auhmani, M. M. Hassani, Quantitative and qualitative analyses of Dimension Reduction Methods effect on the classification of mammographic images, IJCSI International Journal of Computer Science Issues, Volume 11, Issue 6, No 1, November 2014.
[36]N. Hamdi, K. Auhmani, M. M. Hassani, A comparative study of dimension reduction methods combined with wavelet transform applied to the classification of mammographic images, International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No 6, December 2014.