

Clustering Analysis for Extracting Moroccan Health Provinces Typology According to Breast and Cervical Cancer Early Screening

Meryem Chakkouch¹, Merouane Ertel², Aziz Mengad³, Said Amali⁴, Majda Frindy⁵

Informatics and Applications Laboratory (IA)-Faculty of Sciences, Moulay Ismail University, Meknes, Morocco^{1, 2}
Centre for Doctoral Studies "Life and Health Sciences"- Drug Sciences Formation, Laboratory of Pharmacology and Toxicology (LPTR)-Faculty of Medicine and Pharmacy of Rabat (FMPh), Impasse Souissi, Rabat, Morocco³
Informatics and Applications Laboratory (IA), FSJES Moulay Ismail University, Meknes, Morocco⁴
Health Delegation, Ministry of Health and Social Protection, Rabat, Morocco⁵

Abstract—Cancer remains a major global concern, and its screening is a complex public health intervention. In Morocco, breast and cervical cancers are the most frequent malignancies among women, accounting for about half of all diagnosed cases. However, screening participation and coverage still vary across provinces. This study proposes a provincial typology of early screening performance using collected indicators for breast and cervical cancer. Before clustering, we applied several dimensionality reduction methods to improve cluster separability. We adopt a comparative framework that evaluates combinations of DR techniques (PCA, ICA, kernel PCA, t-SNE, and LLE) and clustering algorithms (ACH, K-Means, and GMM) to identify the optimal model with the help of internal validation measures. Kernel PCA with K-Means presents the most optimal model, producing the most coherent province clustering from all tested combinations (DR & algorithm clustering). It demonstrates the best overall separation and compactness according to the evaluation metrics. Three clusters were obtained describing a gradient of early screening system performance: the first group of provinces shows higher screening coverage and stronger diagnostic and referral capacity, the second group demonstrates intermediate performance and differentiated service delivery, and the third group of provinces with low coverage and restrictive access reflects geographic remoteness and service constraints. These results emphasize marked spatial disparity in preventive service performance. They demonstrate how unsupervised learning can support territorial health analysis. The resultant typology can inform targeted action: maintaining and sustaining quality in high-performing provinces, strengthening operations in intermediate-performing provinces, and giving priority to catch-up interventions in low-performing areas.

Keywords—Clustering; PCA; ICA; KPCA; t-SNE; LLE; K-Means; ACH; GMM; breast and cervical cancers early screening

I. INTRODUCTION

Cancer is now one of the major threats to global public health, due to a continued increase in morbidity and mortality. According to the International Agency for Research on Cancer (IARC), approximately 20 million new cases were expected in 2022, with age-standardized incidence rates of 212.6 per 100,000 in men and 186.3 per 100,000 in women. In the same year, nearly 10 million deaths were attributed to cancer, corresponding to age-standardized mortality rates of 109.8 per

100,000 in men and 76.9 per 100,000 in women. Strikingly, about two-thirds of these deaths occur in low- and middle-income countries, where mortality levels approach the incidence rates [1].

Early detection initiatives play an essential role in reducing the incidence of certain cancers. Through the implementation of effective awareness campaigns, timely cancer screening programs, and rapid therapeutic interventions, substantial decreases in cancer-related morbidity and mortality can be achieved [2],[3],[4],[5]. In line with WHO reports, between 30% and 50% of cancer cases are preventable through the control of risk factors or the implementation of screening strategies [6]. The burden of these cancers can be reduced through the adoption of early detection programs followed by appropriate treatment [7].

Morocco's health agenda has prioritized the early screening of breast and cervical cancer, as these two cancers represent the most commonly diagnosed cancers among women, accounting for 38.1% and 8.1% of cases, respectively, according to cancer registry data. All levels of the Moroccan healthcare system integrate all activities related to the early detection of breast and cervical cancer.

Studies related to breast and cervical cancer screening reveal geographic disparities in the availability of screening services and participation rates, which require the implementation of analytical methods that can characterize regional performance. Geospatial techniques have proven effectiveness in mapping these variations, where it is identifying persistent clusters of high and low screening engagement that are usually associated with urban-rural divides, which emphasizes that geographic location is a decisive factor in the effectiveness of healthcare services. [8],[9],[10],[11]. Unsupervised machine learning techniques are also used for regrouping regions based on multidimensional data on screening and patient demographics, to identify distinct clusters and typologies that can support and guide public health strategies and adapt resource allocation [12].

In addition to these spatial analyses, Sassenou et al. [13] applied unsupervised ML on longitudinal data to define three cervical cancer profiles. They revealed the ability of typological

models to expose structural barriers and guide strategies that are aimed at improving screening.

The main objective of this article is to apply clustering techniques to find typologies of Morocco's health provinces and split them into distinct categories based on indicators of early detection of breast and cervical cancer. This is accomplished through different dimensionality reduction techniques and clustering algorithms. This approach aims to derive a reproducible, policy-relevant typology of provinces that reflects differences in early screening of breast and cervical cancer. In other terms, the typology reached, through this analysis, can be used to help in decision-making regarding the implementation of the early screening program for these two cancers at the provincial level and to guide future actions so as to be adapted to each category of typology.

We compare multiple embedding methods (PCA, t-SNE, ICA, LLE, and KPCA) with several clustering algorithms (k-means, agglomerative/hierarchical, Gaussian mixture models) and select models using internal validity indices (Silhouette, Calinski–Harabasz, Davies–Bouldin). Finally, we profile clusters against contextual variables (facility density, urban–

rural mix, population size) and translate findings into actionable program recommendations.

The article has three main parts. The first part, Materials and methods, includes data collection, data preprocessing, dimensionality reduction techniques, machine learning algorithms, and performance indicators. The second part, Results and Discussion, includes descriptive statistics, clustering outcomes, performance comparison, and sanitary province typology. The third part, Conclusion, summarizes the main findings and explains the importance and the implications of the study.

II. PROPOSED MODEL

In this article, we propose to identify the Moroccan health provinces typology based on a classic clustering methodology. We start by describing the dataset (data collected, observations, and features) and the preprocessing procedures (scaling). Then we opt for a comparative analysis of different models obtained from the interaction between the DR method and clustering algorithm. Finally, we interpret results and discuss their practical implications for public health decisions and resource allocation (Fig. 1).

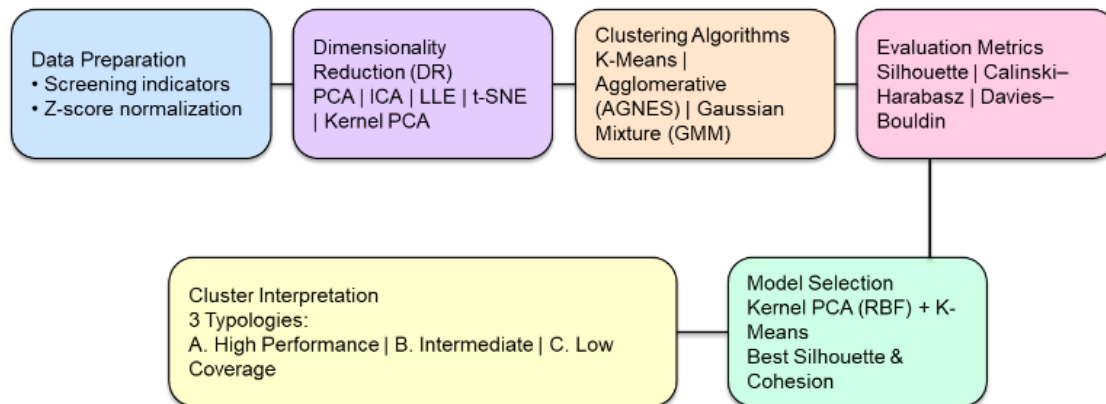


Fig. 1. Clustering methodology.

III. MATERIALS AND METHODS

A. Data Collection

Morocco undertakes the implementation of an early detection program for breast and cervical cancers and its integration into the reproductive health activity package delivered in primary healthcare facilities, since 2010, characterized by the launching of the National Cancer Prevention and Control Plan. This early detection program targets women aged 45 to 69 for breast cancer early detection and women aged 30 to 49 for cervical cancer early screening, and it is based on clinical examination for breast cancer and on the method of visual inspection of the cervix with acetic acid (VIA).

Data related to this program is obtained from the Health Information System for Reproductive Health, Child Health, and Curative Care (SREC). SREC includes important data on early detection programs, such as the number of women screened for the first time, follow-up visits, referrals for diagnostic assessment, imaging, and biopsy procedures, and confirmed

cases. For cervical cancer, the system also reports results from VIA, colposcopy, detection of precancerous lesions, and invasive cancer diagnoses. In addition, SREC includes information on health system resources, including the number of facilities, physicians, nurses, and public health centers.

The dataset is publicly available from the document "Health in Figures" [14]. It is an annual report summarizing statistical information and data related to the national healthcare system. It highlights the availability of both public and private healthcare services, as well as the activities and outputs of public healthcare institutions at national, regional, and local (provincial/prefectural) levels. This document is considered an official data source provided by the Ministry of Health and Social Protection. It serves as an important reference for healthcare administrators, medical professionals, researchers, students, and other individuals interested in health-related data [14].

In this study, we collected 19 features from 82 provinces/prefectures. These features are organized into the following domains (Table I):

TABLE I. LIST OF STUDY VARIABLES

Breast Cancer	Number of women examined for the first time for breast cancer
	Number of women returning for breast cancer follow-up
	Number of women referred to the CRSR for breast cancer
	Number of women received for early diagnosis of breast cancer
	Mammograms performed for breast cancer
	Ultrasounds performed for breast cancer
	Biopsies requested for breast cancer
	Confirmed breast cancer cases
Cervical Cancer	Number of women undergoing VIA (Visual Inspection with Acetic Acid) for the first time
	Number of women returning for cervical cancer follow-up
	Number of women referred to the CRSR for positive VIA results
	Number of women received for early diagnosis of cervical cancer
	Colposcopies performed for cervical cancer
	Number of precancerous cervical lesions detected
	Number of invasive cervical cancer cases diagnosed
Healthcare Supply (2022)	Number of Primary Health Facilities (ESSP)
	Total number of public physicians
	Total number of nurses
	Public hospital beds

B. Data Preprocessing

Every machine learning journey starts with data preprocessing. Feature scaling is one of the most important steps in preprocessing.

When working with datasets that include variables measured in different units or scales (Number of women examined for the first time for breast cancer vs. confirmed breast cancer cases), directly combining them in calculations can lead to skewed results. Variables with larger numerical ranges (like Number of women examined for the first time for breast cancer) tend to overpower those with smaller ranges (like Confirmed breast cancer cases), even if both are equally important [15].

So, we first divide features by the population, and then we standardize the resulting per capita values.

As the features are numbers, and as it is necessary to take into consideration the demographic weight of each province to allow for fair comparisons across provinces, we divided all variables for each province by the size of the target population.

To normalize the data, we use Z-score standardization, a technique that centers features around a mean of 0 and scales them to a standard deviation of 1. This involves two steps:

- Subtracting the mean value of each feature from its data points.
- Dividing the outcome by the feature's standard deviation.

The final standardized values allow the creation of a unitless comparable scale between the different variables, representing

the number of standard deviations that separate each observation from the mean.

Given an X feature, the formula of the new feature standardized X' is defined as: $X' = \frac{X - \mu}{\sigma}$; μ and σ are the sample mean and standard deviation of the feature X .

C. Feature Engineering: Dimensionality Reduction Techniques

We will apply 5 different techniques of dimensionality reduction: Principal Component Analysis (PCA), Gaussian kernel Principal Component Analysis (Gaussian kPCA), Independent Component Analysis (ICA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Locally Linear Embedding (LLE).

1) *Principal Component Analysis (PCA)*: PCA is a linear dimensionality-reduction technique that compresses high-dimensional data into a smaller set of uncorrelated variables, the principal components, while preserving most of the original variance. Mathematically, PCA solves an eigenvalue problem on the covariance (or correlation) matrix of mean-centred, usually standardized, features, yielding an orthogonal basis ordered by descending variance. Equivalently, PCA can be performed via the Singular Value Decomposition (SVD) of the data matrix, which provides the same principal directions and variances in a numerically stable way [15],[18].

2) *t-Distributed Stochastic Neighbor Embedding (t-SNE)*: t-SNE is a non-linear dimensionality reduction technique designed for visualizing high-dimensional data in 2D or 3D while preserving local structures (clusters). It uses Gaussian conditional probabilities to model pairwise similarities in high-dimensional space:

$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)} \quad (1)$$

with $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$, and in low-dimensional space using a Student t-distribution

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} \exp\left(-\frac{\|y_i - y_k\|^2}{2\sigma_i^2}\right)} \quad (2)$$

The algorithm minimizes the Kullback-Leibler (KL) divergence $KL(Y) = \sum_{i \neq j} p_{ij} \log(p_{ij}/q_{ij})$ through a gradient descent, prioritizing the preservation of local neighbors.

Unlike linear methods (PCA and ICA), t-SNE focuses on maintaining the relative distances between nearby points, making it ideal for uncovering patterns in complex datasets (images, gene expression, ...). [19].

3) *Gaussian kernel Principal Component Analysis (Gaussian kPCA)*: kPCA is a non-linear extension of PCA that leverages kernel methods to capture complex, non-linear patterns in data. The most widely used kernel is the Gaussian radial-basis function (RBF): $k_\gamma(x, x') = \exp(-\gamma \|x - x'\|^2)$; $\gamma = \frac{1}{2\sigma^2} > 0$ & x and x' are two data samples (row vectors) from the original input space R^p [20],[22].

4) *Independent Component Analysis (ICA)*: ICA is a statistical technique used to separate a multivariate signal into additive, statistically independent components. Unlike PCA (which finds orthogonal directions of maximum variance) or kPCA (which handles non-linearity), ICA focuses on uncovering hidden factors or sources that are independent of each other. It is widely used in blind source separation (BSS) and signal processing [15][17]. ICA models $X = AS$, where S : Matrix of independent source signals and A : Mixing matrix (unknown linear transformation).

5) *Locally Linear Embedding (LLE)*: LLE is a nonlinear dimensionality reduction method that preserves local linear relationships by reconstructing each data point x_i as a weighted combination of its k -nearest neighbours. For high-dimensional data, it first solves $\min_W \sum_i \|x_i - \sum_j W_{ij} x_j\|^2$ subject to $\sum_j W_{ij} = 1$, yielding weights W_{ij} that encode local geometry. It then finds low-dimensional embeddings y_i by minimizing $\min_Y \sum_i \|y_i - \sum_j W_{ij} y_j\|^2$ under the constraint $Y^T Y = I$ (orthonormal coordinates). LLE excels at unravelling manifolds and is deterministic, but struggles with sparse/noisy data, global structure preservation, and tuning the neighbourhood size k .

D. Machine Learning Models

We will apply three clustering algorithms to compare their performance and ability to reveal relevant structures in the data, and subsequently retain the most consistent and interpretable partition. First, we will apply the famous K-means algorithm, then the Gaussian Mixture Model (GMM), and Agglomerative clustering. These three algorithms present three complementary and basic groups of clustering approaches, including centroid-based, distribution/model-based, and minimizing hierarchical variance [24].

1) *K-means*: K-means is one of the most widely used methods for partitioning a dataset into k distinct, non-overlapping clusters. Its goal is to group observations so that points within the same cluster are as similar as possible (intra-cluster cohesion) and points in different clusters are as dissimilar as possible (inter-cluster separation). It aims to minimize the variance within each cluster by iteratively refining centroids (cluster centers) using a distance matrix, such as Euclidean distance. It consists of minimizing the sum of squared errors (SSE) within clusters: $SSE = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$ where C_i : Points in cluster i and μ_i : Centroid of cluster i . [21],[23].

2) *Gaussian Mixture Model (GMM)*: GMM is a probabilistic unsupervised learning algorithm that assumes data points are generated from a mixture of k Gaussian (normal) distributions. Unlike k-means (which uses hard clustering), GMM provides soft clustering by assigning probabilities to each data point for belonging to each cluster.

GMM algorithm constitutes a powerful tool for modeling complex data distributions but requires careful tuning of k and regularization (diagonal covariance) to avoid overfitting [23].

3) *Agglomerative*: AHC is an unsupervised learning algorithm that tries to group data points that are similar into clusters according to their proximity. It is a bottom-up method that creates a hierarchy of clusters by iteratively merging the closest pairs of data points or clusters until all points converge into a single cluster. Starting with each point as its own cluster, it uses linkage criteria, such as average (mean distance), complete (maximum distance), single (minimum pairwise distance), or Ward's method (minimizing merged cluster variance) to compute inter-cluster distances and guide merges. AHC avoids predefining the number of clusters (a key advantage over k-means) and captures hierarchical structures. It produces a dendrogram visualizing data relationships and allowing users to divide it at a desired height to extract clusters. It is recommended to use it for small to medium datasets, like in biology (gene clustering), taxonomy studies, and social network analysis.

E. Performance Indicators

The evaluation of the effectiveness of the clustering results, we will apply three distinct internal validation metrics: the Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Index.

1) *Silhouette score*: The SI is a metric that evaluates clustering quality through intra-class proximity (how closely data points are grouped within their clusters) and inter-class distance (how distinct they are from other clusters):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

where, $a(i)$ is the average intra-cluster distance and $b(i)$ is the average distance to the nearest neighboring cluster, the score ranges from -1 to 1. Values near 1 indicate well-defined clusters, 0 suggests overlapping clusters, and negative values imply misassignment.

2) *Calinski-Harabasz score*: The CHI index is defined as the ratio of the weighted sum of between-cluster variance (separation) to within-cluster variance (compactness) :

$$CH = \frac{SS_B / (k-1)}{SS_W / (n-k)} \quad (4)$$

where,

- $SS_B = \sum_{i=1}^k n_i \cdot \|\mu_i - \mu\|^2$: is the between-cluster dispersion (sum of squared distances between cluster centroids μ_i and the global centroid μ , weighted by cluster size n_i).
- $SS_W = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$ is the within-cluster dispersion (sum of squared distances between points and their cluster centroid).
- k : number of clusters, n : total data points

A higher score indicates better-defined clusters, as it reflects tightly grouped points within clusters and distinct separation between them.

3) *Davies-Bouldin index*: The DBI is calculated as the ratio of the sum of the within-cluster scatter to the between-cluster scatter :

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left(\frac{S_i + S_j}{d(c_i, c_j)} \right) \quad (5)$$

- $S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i\|$: average intra-cluster distance for cluster C_i (compactness).
- $d(c_i, c_j)$: Distance between centroids c_i and c_j (separation).

A small DBI value indicates a compact cluster with tight and well-separated clusters, with 0 being the ideal minimum.

IV. RESULTS AND DISCUSSION

A. Descriptive Statistics

The dataset covers 82 provinces and early-detection activities for breast and cervical cancer in Morocco in 2022, targeting women aged 45–69 for breast screening and 30–49 for cervical screening (nearly 4.8M and 4.6M, respectively). For breast cancer, about 622,000 women were screened for the first time, and 185,000 returned for follow-up; 16,000 were referred to CRSRs, 17,000 underwent diagnostic work-up (including 12,000 mammograms, 13,000 ultrasounds, and 2,500 biopsies), resulting in 1,600 confirmed cases. For cervical cancer, VIA reached 155,000 first-time and 47,000 follow-up visits; 7,500 VIA-positive women were referred, 6,000 underwent diagnostic evaluation with 5,800 colposcopies, yielding 700 precancerous lesions and 80 invasive cancers (nearly 12% and 1.4% of colposcopies, respectively). For healthcare supply, in 2022, Morocco accounts for 3015 ESSP, 13 762 public physicians, 37 376 nurses, and 27 401 hospital beds. The distribution of these variables presents inequities at the regional, provincial, and municipal levels.

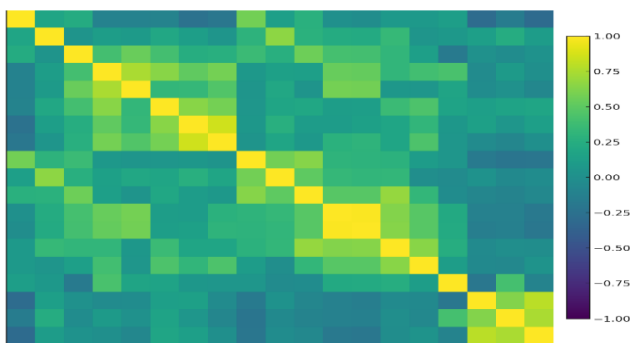


Fig. 2. Correlation plot.

Fig. 1 represents the correlation heatmap, which visualizes the existence or nonexistence of a linear relation between two variables. It presents a clear block structure: a breast pathway block (first visit, return, referral, imaging, biopsy, confirmed), a cervical pathway block (VIA, returns/referrals, diagnostic visit, colposcopy, lesions/invasive), and a smaller capacity block (ESSP, physicians, nurses, public health facilities). The correlation within the block is high, which motivates z-score standardization and dimensionality reduction before clustering.

B. Clustering Analysis

Before we start the clustering analysis, we must find a good number of clusters. Different methods can be used in this case; in this study, we opted for the elbow technique to choose the optimal number of clusters to retain. This method examines the percentage of variance explained as a function of the number of clusters. Notably, we choose a certain number of clusters so that adding another cluster does not significantly improve the data modelling.

In our case, according to diagram (Fig. 3), the optimal number of clusters to choose is three, where the leap of inertia clearly appears at $k = 3$.

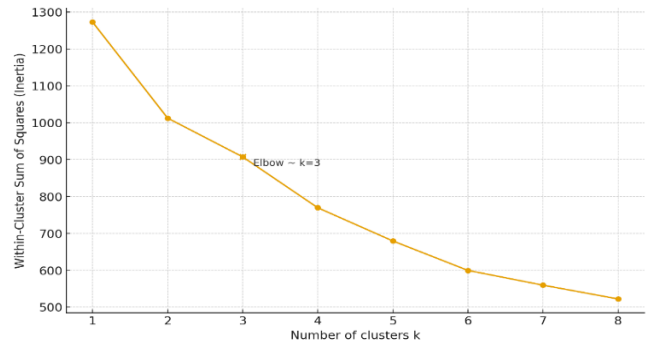


Fig. 3. Optimal number of clusters: elbow method.

Fig. 4 illustrates the interdependence between data representation and clustering methodology. Each row corresponds to a clustering algorithm (K-Means, Gaussian Mixture Model, and Agglomerative Clustering) while each column depicts a distinct two-dimensional embedding of the provincial dataset (Original Space, PCA, KPCA, ICA, LLE, and t-SNE). Each DR method projects the dataset into a two-dimensional subspace, and the three clustering algorithms partitioned the provinces into three distinct clusters.

It shows that the separability and the coherence of the resulting clusters differ according to the selected DR method, which is a basic principle in unsupervised learning. Also, the visualizations of the convex hulls in the different subplots demonstrate how the DR technique modulates the data's geometry and how each clustering algorithm distinctly interprets the emergent structures, which vary in their degree of overlap, curvature, and separation.

C. Performance Evaluation

The performance indicators of the resulting models of clustering are presented in Table II below. They quantify the quality of the 18 clustering models obtained from the combinations of DR techniques and clustering algorithms.

KPCA followed by K-Means achieved the best performance indicators, reaching a high SI amounting to 0.42, a high CHI rising to 67.64, and a low DBI of 0.81, and thus provided the most effective and stable clustering. These measures indicate that the resulting clusters are cohesive, well separated, and have minimal internal variance. The cluster sizes ([21, 30, 16]) are balanced, indicating that the model avoids overfitting small and marginal clusters.

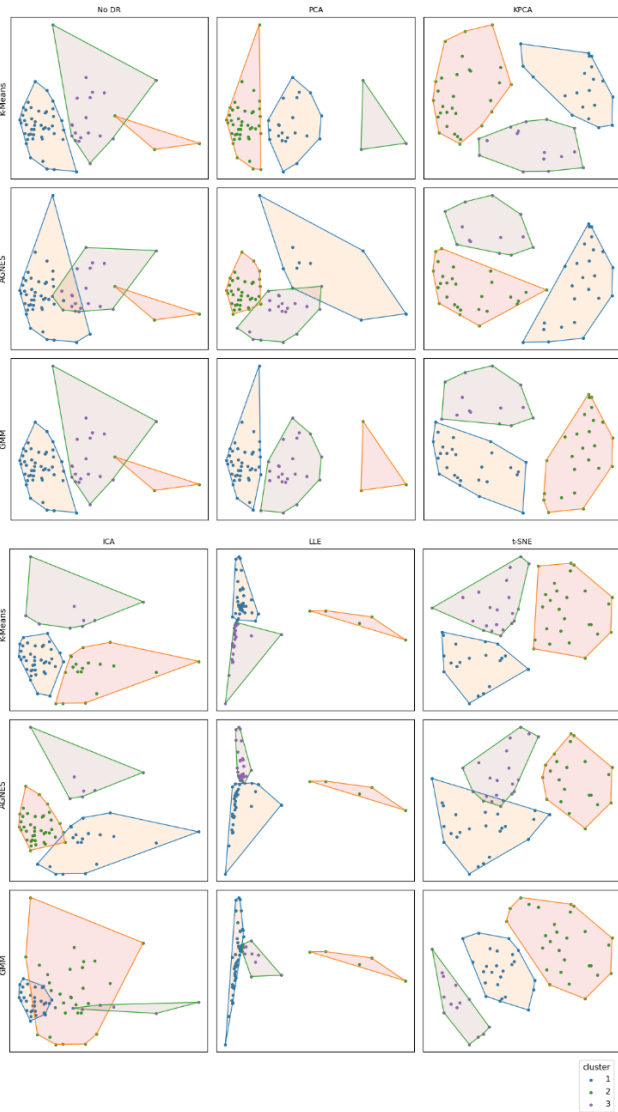


Fig. 4. Two-dimensional visualization of clustering results across dimensionality reduction methods and clustering algorithms.

KPCA’s strength is reflected in its ability to project the data into a higher-dimensional feature space, capturing nonlinear relationships that are invisible to linear methods like PCA or ICA. K-Means, in turn, benefits from this transformation because the non-linear projection reshapes the data into more spherical forms that align with its partitioning assumptions.

The LLE results show the highest numeric indices overall, K-Means reaches an SI of 0.494, a CHI of 80.99, and a DBI of 0.62, but the clustering structure includes a very small cluster of only five provinces ([38, 5, 24]), which explains these internal validation metrics. Indeed, small clusters are automatically compact and far from others, leading to higher scores that do not reflect meaningful separability. GMM’s performance is the least stable, due to its reliance on Gaussian assumptions that rarely match for complex and heterogeneous datasets.

In sum, the non-linear DR techniques surpass linear ones, and the K-Means algorithm benefits more from these

transformations compared to GMM and Agglomerative clustering.

TABLE II. PERFORMANCE INDICATORS

DR Technique	Clustering Algorithm	Silhouette	Calinski-Harabasz	Davies-Bouldin	Cluster Sizes
No DR	<i>K-Means</i>	0,2416 64921	13,527687 3	1,71849 669	[45, 3, 19]
	<i>AGNES</i>	0,1948 55939	12,876148 55	1,70336 1865	[44, 3, 20]
	<i>GMM</i>	0,2416 64921	13,527687 3	1,71849 669	[45, 3, 19]
PCA	<i>K-Means</i>	0,4012 54732	40,118088 32	0,92608 148	[22, 42, 3]
	<i>AGNES</i>	0,3820 36886	35,404338 49	1,08213 5195	[10, 37, 20]
	<i>GMM</i>	0,4010 39442	39,989574 43	0,93727 1776	[41, 3, 23]
KPCA	<i>K-Means</i>	0,4177 81667	67,641716 58	0,811564 069	[21, 31, 15]
	<i>AGNES</i>	0,3974 21428	57,417415 39	0,87850 2454	[26, 28, 13]
	<i>GMM</i>	0,4159 00138	63,545784 66	0,85530 0217	[27, 26, 14]
ICA	<i>K-Means</i>	0,4090 5432	38,853541 62	0,90790 7981	[39, 19, 9]
	<i>AGNES</i>	0,4022 35305	35,816616 68	0,94177 9972	[20, 40, 7]
	<i>GMM</i>	0,2368 90648	14,927386 25	1,56033 8644	[31, 30, 6]
LLE	<i>K-Means</i>	0,4940 05345	80,994758 17	0,62013 9336	[38, 5, 24]
	<i>AGNES</i>	0,4453 15966	75,111679 13	0,66912 3424	[32, 5, 30]
	<i>GMM</i>	- 0,0220 72448	25,898906 67	2,13544 227	[49, 5, 13]
t-SNE	<i>K-Means</i>	0,3693 23403	64,405670 17	0,91038 6817	[20, 26, 21]
	<i>AGNES</i>	0,3427 97369	57,645111 08	0,99574 7264	[24, 22, 21]
	<i>GMM</i>	0,3595 72917	59,727108	0,94195 6303	[25, 28, 14]

D. Sanitary Provinces Typology Exploration

The clustering results obtained from kPCA followed by K-Means delineate three distinct typologies of Moroccan provinces according to the performance of the breast and cervical cancer early detection program. These profiles capture territorial disparities in screening coverage, diagnostic capacity, and care of women, which reflects a non-exhaustive coverage of preventive healthcare for women across the Kingdom. Performance was assessed against the national program objectives, namely: 1) achieving at least 40% annual coverage for breast cancer screening; 2) achieving at least 30% annual coverage for cervical cancer screening; 3) achieving at least 80% return rate among participants in breast cancer screening; 4) achieving at least 80% return rate among participants in cervical cancer screening; 5) diagnosing at least 40% of expected breast cancer cases in women aged 40-69 years; and 6) achieving a 100% treatment rate for precancerous cervical lesions detected within the program[16]. Thus, the three clusters obtained are: High Early Screening Performance, Intermediate Early Screening Systems, and Low Early Screening Coverage.

1) *High early screening performance*: This cluster is characterized by advanced and integrated early screening systems. It regroups provinces that register above-average coverage rates for both VIA and mammography screening. It includes Rabat, Marrakech, and Meknes, which demonstrate superior performance in referral completion, biopsy confirmation, and patient follow-up.

Such results can be explained by the density of healthcare infrastructure, the availability of trained personnel, the presence of regional oncology centers which provide level 3 curative services (chemotherapy, radiotherapy, immunotherapy, surgery...), and the integrated information system which allows for the coordination of health data between the different structures.

It is recommended to consolidate these gains so that these provinces prioritize quality assurance mechanisms, reducing diagnostic lead times and beginning treatment. Sustaining high performance will depend on continuous professional development, preventive maintenance of diagnostic equipment, and regular monitoring of screening quality indicators, including SDG indicators such as the mortality rate attributable to these two cancers among women.

2) *Intermediate early screening systems*: This cluster regroups provinces characterized by moderate early detection performance. It includes Agadir-Ida -Ou-Tanane, Berrechid, El Jadida, Fahs-Anjra, and Kenitra. Early screening activities in these provinces are established but still constrained by gaps in patient tracking, referral coordination, and time to care for women patients. The screening coverage varies depending on the availability of healthcare personnel, outreach campaigns, and external support programs.

These territories correspond to systems where the early detection program for breast and cervical cancer requires additional efforts to improve their performance indicators. To enhance service continuity and program monitoring in this typology of provinces, it is recommended to strengthen referral pathways, introduce digital patient registries, and reinforce data reporting standards.

The Group B provinces constitute the target for strengthening infrastructure, equipment, and human resources in order to achieve the objectives set by the Ministry of Health and Social Protection.

The Group B provinces constitute the target for strengthening infrastructure, equipment, and human resources in order to achieve the objectives set by the Ministry of Health and Social Protection.

3) *Low early screening coverage*: The third cluster includes provinces that experience difficulties in providing early screening services. It comprises Al Haouz, Chichaoua, Boulemane, Chefchaouen, and Berkane, which are predominantly rural and geographically dispersed. Screening coverage remains significantly below the national average, and diagnostic follow-up (ultrasound, biopsy, or CRSR referral) is irregular.

Challenges in these territories stem from shortages of health personnel, insufficient medical equipment, and problems of access and accessibility to referral hospitals. In addition to the annual early detection campaign for breast and cervical cancer (Pink October), screening relies on mobile and sporadic campaigns, which lead to irregular coverage and poor follow-up.

For this group, a comprehensive catch-up strategy is essential. This should include the establishment of permanent VIA units at the primary-care level, quarterly mobile screening caravans, and improved referral and transport mechanisms to regional oncology centers, as well as awareness-raising sessions in maternal education classes.

V. CONCLUSION

The study applies a comparative analysis that integrates a combination of different DR techniques and clustering algorithms. The objective was to retrain the best and most suitable model for partitioning the Moroccan health provinces based on breast and cervical early detection indicators. The model built from integrating kPCA for dimensionality reduction and K-Means for clustering has provided three cluster structures that present a clear gradient of early screening system maturity across the Moroccan Kingdom. It succeeds in identifying reasonable clusters that reflect real territorial differences in health system performance and preventive service delivery.

The first cluster includes provinces with high early screening coverage, a well-defined and well-structured care pathway, typically concentrated in major urban centers and regions with better healthcare resources. The second cluster represents an intermediate group in which early screening remains insufficient due to various organizational challenges and coordination issues. The third cluster concentrates provinces with limited access, weak service integration, and low participation rates, which correspond to geographical isolation and underserved areas.

This clustering highlights that the spatial heterogeneity of preventive and curative healthcare in Morocco depends on socio-geographic determinants such as population density, health workforce distribution, and proximity to oncology reference centers. The analytical framework also demonstrates the added value of machine learning approaches, specifically unsupervised clustering, for uncovering underlying performance patterns that conventional descriptive statistics may obscure.

Finally, this typology provides the Ministry of Health and Social Protection with a decision-support tool. It allows allocating resources strategically, designing tailored capacity-building programs, and implementing monitoring frameworks aligned with the maturity level of each cluster. Sustaining the gains of high-performing regions, consolidating intermediate systems, and accelerating catch-up in low-coverage provinces can contribute to a more equitable and efficient national early screening program.

These insights are particularly timely in the context of Morocco's ongoing health-system reform, framed by Framework Law 06-22 and the extension of compulsory health insurance, which aims to modernize health infrastructure, strengthen human resources, and improve territorial governance

of care. In this regard, the creation of Territorial Health Groupings (GST), designed to pool human and financial resources across a given territory, aims to ensure a more equitable and balanced supply of health services and to reduce health inequalities.

REFERENCES

- [1] F. Bray *et al.*, “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA A Cancer J Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024, doi: 10.3322/caac.21834.
- [2] J. Gutierrez-Cardenas, “Breast Cancer Classification through Transfer Learning with Vision Transformer, PCA, and Machine Learning Models,” *IJACSA*, vol. 15, no. 4, 2024, doi: 10.14569/IJACSA.2024.01504104.
- [3] P. Sasieni, R. Smittenaar, E. Hubbell, J. Broggio, R. D. Neal, and C. Swanton, “Modelled mortality benefits of multi-cancer early detection screening in England,” *Br J Cancer*, vol. 129, no. 1, pp. 72–80, Jul. 2023, doi: 10.1038/s41416-023-02243-9.
- [4] L.-S. Béatrice *et al.*, “Breast-Cancer Screening — Viewpoint of the IARC Working Group,” *n engl j med*, 2015.
- [5] M. Arbyn *et al.*, “Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis,” *The Lancet Global Health*, vol. 8, no. 2, pp. e191–e203, Feb. 2020, doi: 10.1016/s2214-109x(19)30482-6.
- [6] World Health Organization. Preventing cancer. <https://www.who.int/activities/preventing-cancer>.
- [7] World Health Organization. Cancer. <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [8] A. Chandak, P. Nayar, and G. Lin, “Rural-Urban Disparities in Access to Breast Cancer Screening: A Spatial Clustering Analysis,” *The Journal of Rural Health*, vol. 35, no. 2, pp. 229–235, Mar. 2019, doi: 10.1111/jrh.12308.
- [9] J. R. Meliker, G. M. Jacquez, P. Goovaerts, G. Copeland, and M. Yassine, “Spatial cluster analysis of early stage breast cancer: a method for public health practice using cancer registry data,” *Cancer Causes Control*, vol. 20, no. 7, pp. 1061–1069, Sep. 2009, doi: 10.1007/s10552-009-9312-4.
- [10] R. W. Amin, B. A. Fritsch, and J. E. Retzliff, “Spatial Clusters of Breast Cancer Mortality and Incidence in the Contiguous USA: 2000–2014,” *J GEN INTERN MED*, vol. 34, no. 3, pp. 412–419, Mar. 2019, doi: 10.1007/s11606-018-4824-9.
- [11] C. T. Nguyen, I. Song, I. Jung, Y. Choi, and S. Kim, “Changes in spatial clusters of cancer incidence and mortality over 15 years in South Korea: Implication to cancer control,” *Cancer Medicine*, vol. 12, no. 16, pp. 17418–17427, Aug. 2023, doi: 10.1002/cam4.6365.
- [12] H. Bao *et al.*, “How Can a High-Performance Screening Strategy Be Determined for Cervical Cancer Prevention? Evidence From a Hierarchical Clustering Analysis of a Multicentric Clinical Study,” *Front. Oncol.*, vol. 12, p. 816789, Jan. 2022, doi: 10.3389/fonc.2022.816789.
- [13] J. Sassenou, V. Ringa, M. Zins, A. Ozguler, S. Paquet, and L. Rigal, “Underuse, overuse, and guideline-based use of cervical cancer screening: social disparities in temporal screening trajectories in the French CONSTANCES cohort,” *BMC Women’s Health*, vol. 25, no. 1, p. 495, Oct. 2025, doi: 10.1186/s12905-025-03966-y.
- [14] MSPS, Santé en chiffre 2022
- [15] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. New internet. edition. Always Learning. Harlow: Pearson, 2014
- [16] National Cancer Prevention and Control Plan 2020-2029
- [17] X. Wang *et al.*, “Integrated Metabolomics-KPCA-Machine Learning framework: a solution for geographical traceability of Chinese Jujube,” *Food Chemistry: X*, vol. 31, p. 103069, Oct. 2025, doi: 10.1016/j.fochx.2025.103069.
- [18] Xing Xiaoxue, Liu Fu, Shang Weiwei, Li Wenwen, and Zhang Yu, “Research of PCA and KPCA in the characteristics simplicity of the gene data,” in *Proceedings of 2013 2nd International Conference on Measurement, Information and Control*, Harbin: IEEE, Aug. 2013, pp. 669–672. doi: 10.1109/MIC.2013.6758051.
- [19] V. Bakiasi, M. Muça, and R. Kapçiu, “Dimensionality Reduction: A Comparative Review using RBM, KPCA, and t-SNE for Micro-Expressions Recognition,” *IJACSA*, vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.0150135.
- [20] J. Yang, L. Peng, L. Luo, W. Li, and Y. Chen, “Wind power forecasting using hybrid ICEEMDAN-KPCA and IWOA-BiLSTM models,” *International Journal of Electrical Power & Energy Systems*, vol. 173, p. 111445, Dec. 2025, doi: 10.1016/j.ijepes.2025.111445.
- [21] M. Faizan, M. F., S. Ismail, and S. Sultan, “Applications of Clustering Techniques in Data Mining: A Comparative Study,” *IJACSA*, vol. 11, no. 12, 2020, doi: 10.14569/IJACSA.2020.0111218.
- [22] L. Liu and H. Shao, “Study on neutron–gamma discrimination method based on the KPCA-GMM,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1056, p. 168604, Nov. 2023, doi: 10.1016/j.nima.2023.168604.
- [23] T. T. Tin *et al.*, “Natural Disaster Clustering Using K-Means, DBSCAN, SOM, GMM, and Mean Shift: An Analysis of Fema Disaster Statistics,” *ijacsa*, vol. 15, no. 9, 2024, doi: 10.14569/IJACSA.2024.0150968.
- [24] C. X. Gao *et al.*, “An overview of clustering methods with guidelines for application in mental health research,” *Psychiatry Research*, vol. 327, p. 115265, Sep. 2023, doi: 10.1016/j.psychres.2023.115265.