*School of Electrical & Computer Engineering*

# MASTER THESIS

## Self-Organized Clustering of Big Data: Application on the extraction of cervical classes from backscattering curves

**IOANNA-THEONI VOURLAKI**

Technical University of Crete

***Examination committee***

1. ***Professor Michalis Zervakis***
2. ***Professor Costas Balas***
3. ***Associate Professor Michail G. Lagoudakis***

Table of Contents

# ABSTRACT

Data mining is an interdisciplinary subfield of computer science. It forms the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and data systems. Machine learning goes often in parallel with data mining, with the first being a supervised scheme whereas the latter focuses more on exploratory data analysis and is known as unsupervised learning. Clustering constitutes an unsupervised learning approach aiming to organize the available data into compact classes according to some notion of similarity. The contribution of clustering in medicine and biology is highly significant.

In this sense, this Master's thesis examines three fundamental aspects of clustering, namely stability, generalizability and separability in order to discover and interpret the appropriate information from the processed data and make clustering attractive for the effective organization of large datasets.

First, (in association with stability problems), we propose a novel algorithmic approach for extracting adequate information from the input dataset and self-organizing data expanding clustering with resampling. This approach aims to derive stable and representative class centers through permutations in the initialization process implemented via the concept of data resampling. Thus, using data resampling with replacement, we produce multiple partitions from k-means based on multiple reruns of the same population. In our approach, the large number of class centroids is then reorganized into tight groups through the mean-shift approach, which rigorously searches for maxima into this new distribution space of meta-data (class centroids).

Then, by further exploring clustering stability problems, we attempt to refine and improve the clustering result by sequentially updating (instead of replacing) centers on the basis of their present and previous positions. Based on this updating strategy, we can exploit both prior expert knowledge and posterior data information from the statistical distribution of the examined population. In this part, we examine the stability problem of k-means, by proposing a novel algorithmic scheme for self-organizing data, adopting a recursive-mode k-means clustering approach.

Thirdly, we examine issues of k-means clustering associated with its generalization ability in organizing big datasets. For this purpose, we exploit a data bootstrapping strategy without replacement. With the generation of multiple datasets of rather small size, we attempt to cover the entire data distribution space and capture its structural properties within the multiple classes generated. Each bootstrap stage exploits the stabilization process of the k-means algorithm. Finally, all class centroids generated from the bootstrap process are considered as (meta-data) samples of higher abstraction, which are organized into classes via the mean-shift approach, similar to the stabilization process.

In association with data re-sampling strategies, we also consider the appropriate use of distance metrics addressing another major problem of data exploratory schemes. We apply our algorithmic developments on data expressing the temporal course of tissue reflection under a specific wavelength. The process of aceto-whitening is of paramount importance in cervical

cancer diagnosis and we examine clustering methodologies for extracting, processing and interpreting the relevant information from the available data. As in most time-series formulations, the response curves considered are characterized by both overall amplitude (or power) characteristics and local shape formations. The proposed metric attempts to capture both of these aspects into a single configuration, which can be parametrically adjusted to the particular application domain.

Overall, the test results indicate the importance of data resampling (and bootstrapping) in the appropriate partitioning of large datasets and the efficient operation of data mining (and clustering) schemes.

# ΠΕΡΙΛΗΨΗ

Η εξόρυξη δεδομένων είναι ένα διεπιστημονικό πεδίο της επιστήμης των υπολογιστών. Είναι η υπολογιστική διαδικασία ανακάλυψης προτύπων σε μεγάλα σύνολα δεδομένων και περιλαμβάνει μεθόδους στη διεπαφή της τεχνητής νοημοσύνης, μηχανικής μάθησης, στατιστικής και συστημάτων ανάλυσης δεδομένων.

Στα πλαίσια αυτά, η ομαδοποίηση αποτελεί μια μη επιβλεπόμενη προσέγγιση μάθησης με στόχο να οργανώσει τα διαθέσιμα δεδομένα σε συμπαγείς κλάσεις σύμφωνα με κάποιο κριτήριο ομοιότητας. Η συμβολή της ομαδοποίησης στην ιατρική και στη βιολογία είναι πολύ σημαντική. Πιο συγκεκριμένα, στην διάγνωση καρκίνου, η μεθοδολογία της εξαγωγής, επεξεργασίας και ερμηνείας των σχετικών πληροφοριών από τα διαθέσιμα δεδομένα είναι υψίστης σημασίας. Με αυτήν την έννοια η παρούσα εργασία εξετάζει τις τρεις βασικές πτυχές της ομαδοποίησης, την σταθερότητα, την γενίκευση και την διαχωρισημότητα των κλάσεων. Στόχος της εργασίας είναι να αξιολογήσει και να ερμηνεύσει τις κατάλληλες πληροφορίες από τα επεξεργασμένα σύνολα δεδομένων και ιδιαίτερα να κάνει την διαδικασία ομαδοποίησης αυτοματοποιημένη και αποδοτική για την οργάνωση μεγάλου όγκου δεδομένων.

Σχετικά με τα προβλήματα σταθερότητας των αποτελεσμάτων επιχειρούμε να βελτιώσουμε το αποτέλεσμα της ομαδοποίησης αναβαθμίζοντας διαδοχικά τα κέντρα των κλάσεων στη βάση της παρούσας και προηγούμενης θέσης τους. Επίσης προσπαθούμε να αξιοποιήσουμε την αρχική κλινική πληροφορία, την μετέπειτα πληροφορία που εξάγεται από την ανάλυση των δεδομένων και την στατιστική κατανομή του εξεταζόμενου πληθυσμού. Υπό το πρίσμα αυτό, εξετάζουμε το πρόβλημα σταθερότητας του k-means αλγορίθμου, προτείνοντας ένα νέο αλγοριθμικό τρόπο για την αυτό-οργάνωση των δεδομένων, υιοθετώντας μία επαναληπτική προσέγγιση στον αλγόριθμο ομαδοποίησης k-means.

Βασιζόμενοι στην περεταίρω διερεύνηση των προβλημάτων σταθερότητας της ομαδοποίησης, προτείνουμε μια νέα αλγοριθμική προσέγγιση για την εξαγωγή πληροφοριών από το σύνολο των δεδομένων εισόδου και την αυτό-οργάνωση των δεδομένων, συνδυάζοντας την ομαδοποίηση με την επαναληπτική δειγματοληψία των δεδομένων. Χρησιμοποιώντας ανα-δειγματοληψία με αντικατάσταση επιχειρούμε να παράγουμε πολλαπλές εκδοχές του ίδιου πληθυσμού και πολλαπλά κέντρα από τον k-means. Έπειτα, ο μεγάλος αριθμός κέντρων που παράγεται, αναδιοργανώνεται σε ομάδες «μετα-δεδομένων», με τη χρήση της προσέγγισης του Mean Shift αλγορίθμου, ο οποίος ψάχνει για τα μέγιστα στο νέο χώρο κατανομής αυτών των μετα-δεδομένων.

Σε ένα τρίτο στάδιο ανάπτυξης, εξετάζουμε την ικανότητα γενίκευσης της ομαδοποίησης k-means για την οργάνωση των μεγάλων συνόλων δεδομένων. Στοχεύοντας στην επίλυση του παραπάνω προβλήματος, εκμεταλλευόμαστε την στρατηγική της επαναληπτικής αναδιάρθρωσης (bootstrapping) δεδομένων χωρίς αντικατάσταση. Πιο συγκεκριμένα, με τη δημιουργία πολλαπλών συνόλων δεδομένων μικρού μεγέθους προσπαθούμε να καλύψουμε το σύνολο του χώρου κατανομής, χαρακτηρίζοντας τις δομικές ιδιότητες των δεδομένων μέσω των πολλαπλών κλάσεων που δημιουργούνται. Κάθε bootstrap στάδιο εκμεταλλεύεται τη

διαδικασία σταθεροποίησης του αλγορίθμου k-means. Τα πολλά παραγόμενα κέντρα οργανώνονται σε κλάσεις μέσω της προσέγγισης του Mean Shift.

Τέλος, σε συνδυασμό με τις προηγούμενες προσεγγίσεις ανα-δειγματοληψίας και αναδιάρθρωσης, εξετάζουμε την χρήση του κατάλληλου μέτρου απόστασης με στόχο την αντιμετώπιση και τη διερεύνηση του τρίτου μεγάλου προβλήματος της ομαδοποίησης, την διαχωρισημότητα των τελικών κλάσεων. Οι μέθοδοι της εργασίας αυτής χρησιμοποιούνται στα δεδομένα χρονικής απόκρισης σε φασματική ακτινοβολία ιστών του τραχήλου της μήτρας. Όπως στα περισσότερα δεδομένα χρονοσειρών (time-series), οι καμπύλες των δεδομένων χαρακτηρίζονται από το μέγεθος κα το σχήμα αυτών. Η προτεινόμενη μετρική στοχεύει στο να ενσωματώσει και τις δυο αυτές πτυχές των δεδομένων σε μια ενιαία μορφή, η οποία μπορεί να προσαρμοστεί στο συγκεκριμένο πεδίο εφαρμογής.

## Acknowledgements

*I would like to thank,*

*My thesis supervisor, Professor Michalis Zervakis for his continuous support, guidance and patience through the time I have been working on this thesis. The door to Prof. Zervakis office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this study to be my own work, but steered me in the right direction whenever he thought I needed it with his immense knowledge.*

*Professor Costas Balas, for the opportunity that gave me to work and develop my thesis based on his data and the valuable contribution as member of the thesis committee.*

*Associate Professor Michail Lagoudakis, for the precious support and confluence as member of the thesis committee.*

*George Livanos, for his valued cooperation.*

*I would like to express appreciation to my fiancé Nikos, for his unconditional and continuous encouragement to strive towards my goal. You are always there for me.*

*Last ,but not least, I must express my very profound gratitude to my parents, my brother and my grand-father for providing me with unfailing support throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.*

# 1. INTRODUCTION

## 1.1 Aspects of Exploratory Clustering & Areas of Contributions

In this thesis, we explore the three aspects of exploratory clustering, namely stability, generalizability and separability, in order to make it attractive for the effective organization of large data. In the first one (associated with stability problems) we propose a novel algorithmic approach for extracting adequate information from the input dataset and self-organize them through the fusing clustering and data resamples. This approach, aims to derive stable and representative class centers through permutations in the initialization process through the concept of data resampling. Through data resampling with replacement, we produce multiple versions of the same population and multiple partitions from k-means, similar to [11]. In our approach, the large number of class centroids is reorganized into tight groups through the MSH approach, which rigorously searches for maxima into this new distribution space of meta- data. Further exploring clustering stability problems, we attempt to refine and improve the clustering result by sequentially updating centers on the basis of their present and previous positions, exploiting both prior expert knowledge and posterior data information from the statistical distribution of the examined population. In this sense we examine the stability problem of k-means, by proposing a novel algorithmic scheme for self-organizing data, adopting a recursive-mode k-means clustering approach.

The second problem of k-means clustering is associated with its generalization ability in organizing big datasets. For this purpose, we exploit the data bootstrapping strategy without replacement. With the generation of multiple datasets of rather small size, we attempt to cover the entire distribution space and capture its structural properties within the multiple classes generated. Each bootstrap stage exploits the stabilization process of the k-means algorithm. Finally, all class centroids generated from the bootstrap process are considered as (meta-) samples of higher abstraction, which are organized into classes via the MSH approach, similar to the stabilization process. In association with the data re-sampling strategies, we also consider the appropriate use of distance metrics addressing the third major problem of data exploratory schemes. As in most time-series formulations, AW curves are characterized by both overall amplitude (or power) characteristics and local shape formations. The proposed metric attempts to capture both aspects into a single configuration, which can be parametrically adjusted to the particular application domain.

Our proposed approach is evaluated on two groups of AW data. The first set is composed of 497 reference samples experimentally labeled by a medical expert based on biopsy. This set is used for assessing stabilization issues and the effectiveness of data resampling in producing meaningful clusters. Towards this direction, the confusion matrix (using the known labels) is used as a means of evaluating the efficiency of the recovered class centers in representing the classes of interest. This stabilization procedure is performed initially using the Euclidean distance as a data separation measure and is repeated using the new combined distance metric. The

comparison of the resulting confusion matrices highlights the potential of the new metric in time-series exploratory analysis.

Testing the generalization ability of our approach, we attempt to classify a set of 100.000 samples of unknown nature. The process of data bootstrapping is implemented many times in order to produce multiple sub-samples from the original population. The two-stage process with the proposed combined distance, i.e. the initial process of k-means on the bootstrap data and the second stage of mean-shift classification of the meta-data, produces the centroids that are finally used to cluster the entire data population. The evaluation of this (generalization) process is only performed on a qualitative manner on the graphically depicted curves of each cluster. Nevertheless, through comparisons with a clustering that utilizes the clinically approved centroids of pathological stages of interest, it is concluded that the new scheme not only achieves better compactness of clusters but also enables the derivation of new classes emerging from the data, which may be worth of further clinical assessment.

The theoretical background beyond clustering and bootstrapping are presented in Chapter 2 and Chapter 3 respectively. The experimental procedure for obtaining the input dataset along with the proposed algorithmic framework is analyzed in Chapter 4, where a sequential application of processing scenarios takes place until the whole information for the classification is derived. Moreover, the results of our methodology are revealed in the same Chapter 5, along with the outcome on stability and generalizability of the current study.

## 1.2 Thesis Concepts

Apart from the selection of the imaging modality for the assessments of cervical cancer, the methodology for extracting, processing and interpreting the relevant information from the available data is of paramount importance. In essence, the examined cases are represented by feature vectors reflecting the important properties of the tissue under examination or the characteristic structures from its imaging methodology. In our case, data vectors reflect the AW course over time for each pixel of the multispectral image in one specific wavelength range that best reflects the cellular deformations of cancerous tissue [16].

One major goal of our study is to explore ways of organizing this data into meaningful classes acceptable by clinicians. More specifically our study aims at characterizing each and every pixel of the recorded sequence of images over time, thus providing quantifiable measures for the state of the lesion and its borders. The clinician's knowledge can be exploited in the beginning of the process by influencing the characteristic distributions of intended pathology classes, or at the end of clustering as a means of evaluating the quality and clinical value of automatically separated groups. The latter scheme is purely based on the data properties and forms the basis of clustering used in an unsupervised form in order to organize the available data into classes according notion of similarity.

The concept behind clustering is that data similarity is enough to describe compact classes in a feature space, with no other kind of information available (e.g. data labels). An additional benefit of such self-organization is the potential of recovering peculiar disease stated expressed

through novel data formations. Nevertheless, the guidance totally by data renders the clustering process immune to peculiarities of the data and the data handling scheme, and results in drastic increase of complexity with the data size as the problem is not NP complete. Over the years, although numerous clustering algorithms have been proposed, k-means approach still sustains its position as one of the most competitive algorithms for clustering. Producing reasonably good results, this time-efficient technique is easily combined with other methods in larger systems aiming at automatically partitioning the entire data set into a set number of $k$ groups. The algorithm operates by initially k random as cluster centers and then iteratively refining them based on the data samples in the average associated with each cluster in the sense of closer distance. The basic limitation of k-means is need of a priori knowledge on the number of candidate classes and the shape of their distribution. Such problems can be overcome utilizing the Mean-Shift (MSH) clustering algorithm modifies by Cheng [17] and Comaniciu [18].

In essence, MSH is an iterative mode detection algorithm in the density distribution space based on moving to a kernel-weighted average of the observations within a smoothing window. This computation is repeated until convergence and is obtained at a local density mode. The MSH approach, even though it is quite effective, is associated with high computational cost, which is dramatically increased with the population size and allows for low parallelization capacity, since it updates class centers based on the entire population.

In the case of big data analysis, it is desirable to preserve the simple and fast structure of the k-means approach, while attempting to overcome its inherent inefficiencies. K-means clustering aims at self-organization of a population based on similarity of data vectors within each class or with a single reference vector for each class. We can identify three types of problems associates with exploratory approaches such as the k-means algorithms. The first relates to the need of partial knowledge regarding the data framework, such as the number of classes. The second issue associates with the stability of algorithms and the influence of initialization to the solution. As described in the sequel, data resampling with different initialization points can address this problem. The third issue relates to the generalization ability of exploratory algorithms designed from limited subsets of the data. It emerges from the need of accommodating new (independent) sample with the same rules and assumptions utilized in the design of the algorithm. This problem can be addressed through data bootstrapping techniques aiming to efficiently sample the entire data distribution space with limited bootstrap datasets and use them to improve the training efficiency and enhance the generalization capacity of the designed algorithm.

Towards the stabilization of exploratory clustering, Lisboa et.al. [19] utilized a data resampling process along with the k-means approach as to explore the meaningful assumption that the partition producing the lowest value of (Euclidean) distance is reproducible index repeated initializations and approximates the optimal estimate of cluster configuration. Lisboa and his team shows that generalization is valid for small number of classes, while in practical application concerning large and big datasets and/ or k takes considerable values, similarity indices close to each other may correspond to totally different cluster partitions. Thus an objective methodology must be adopted in order to generate a single after repeated runs of the same, entire procedure and close to the original data formulation. The authors in [19] proposed

a k-means clustering scheme, selecting a partition of the data into non-overlapping, stabilized subsets with the number k of clusters being calculated through repeated application of the standard k-means algorithm. This sampling of a single partition is controlled by combining two performance metrics, one that guides the intra-cluster separation and one that evaluates the inter-cluster stability. The development of this iterative process is to sample the original dataset, randomly initialize this samples set, select a fraction of the results that produce best intra-cluster variation, calculate the intra-cluster distances for all possible pairs, return their median value of the median intra-cluster index. The issue of k-means stabilization via random data initializations has been further extended in [20] on structured vector data representing the temporal curves of cervical tissues AW response.

Repeated iterations of the clustering algorithm are often used, starting from different initial points as a means to derive many ''possible'' partitions, which can then be used to formulate a distribution for inferring the most likely partition. In this form, the stabilization approach shares concepts with data permutation in the generation of population statistics. Furthermore, data bootstrapping through resampling from the original population is used (also in iterative fashion) for generalizing the distribution patterns estimated from a limited dataset or the predictive performance of statistical schemes based on this dataset. It aims to produce many representative data sub-populations, partially capturing all structural details of the original data distribution and, in this form, simulate on the average the probability distribution function that can give rise to any new dataset obeying the structure of the original population. Bootstrapping strategies play an important role in resolving classification-based issues associated with uncertainly and generalizability.

Bootstrapping constitutes a data-based simulation scheme, aiming at finding estimators of the parameters if interest along with confidence intervals. It attempts to estimate the sampling distribution via resampling (with or without replacement) from the original sample dataset, merely based on the assumption that the sample set is a good representation of the unknown population. Bootstrap distributions usually approximate he shape, spread, and bias of the actual sampling distribution and are centered at the value of the statistic form the original data, while the sampling distribution is centered at the value of the parameter in the population, plus any bias [21].

Besides the need to improve the performance of the used algorithm, a serious need in exploratory data analysis is the consideration of the algorithmic similarity itself, which is vital in the derivation of representative class centers. Even though class organization usually proceeds with Euclidean distance, other strategies and/or distance metrics may be more appropriate under general data distributions. The case of robust distance metrics has been investigated in [22, 23] without, however, significant improvement on class discrimination. Towards another direction, [24] explores alternatives of k-means algorithm with the k-intervals formulation, which is based on intervals instead of centers. This consideration can alleviate the effects of data outliers and provide efficient and robust clusters without well-defined class centroids. Considering the above implications, the summarization of a class and its individual characteristics by a single point in the data space (cluster center) needs particular attention, since it might result in loss of key information regarding the sample population. For this reason, we explore combined distance

measures influencing the separability of classes and guiding the definition of class centers, which reflect complementary data characteristics from diverse viewpoints. More specifically, we develop a new distance measure appropriate for dynamic time-series characterization, which combines i) amplitude and ii) shape differences.

## 1.3 State of the art on related clustering applications

During the last decade, data mining has emerged as a growing interdisciplinary field that merges together database, statistics, machine learning and related areas in order to extract useful knowledge from data. Clustering is one of the fundamental operations in data mining. This section studies addressing applications of clustering in disease diagnosis with three of the most often used techniques, namely k-means, self-organizing map and mean shift. K-means is used widely in many applications and a variety of field in order to organize the data into meaningful classes. Many alternations and applications of k-means operation have been proposed and published.

K-means is a valuable characterization tool contributing in the diagnosis of several diseases detection. This algorithm was used to automatically detect of erytho-squamous diseases [25]. K-means was utilized to find the five erythomato-squamous diseases and the total classification accuracy of k-means clustering was 94.22%.

Clustering schemes have also been applied in the staging of brain tumors, especially for high-grade gliomas (HGGs) which include glioblastoma (GBM) and anaplastic astrocytoma (AA) and are the most ordinary substantial brain tumors in adults. In [26], previously unidentified prognostic subclasses of high-grade astrocytoma are discovered to resemble stages in neurogenesis. Poor prognosis subclasses demonstrate markers of proliferation or of angiogenesis. In this sense there is a wide interest to define markers for each of three subclasses using k-means clustering to assign tumors to subclass. Based on microarray analysis, k-means defined HGG subclasses designated as proneural, proliferative and mesenchymal in order to identify the dominant features from the gene list that characterizes each subclass.

The case of multiple myeloma (MM), a malignancy of terminally differentiated plasma cells homing and expanding in the bone marrow, is characterized by a tremendous heterogeneity in outcome following standard and high-dose therapies. With the aim of molecularly defining high-risk disease, [27] presented a microarray analysis on tumor cells from 532 newly diagnosed patients with multiple myeloma (MM) treated on two separate protocols. K-means is performed to separate the small right-hand mode from the largest distribution. This algorithm is applied independently to produce an independent cutoff for high versus low gene expressions.

Plant diseases have turned into problems of high importance for the countries that depend on agriculture as a basis of economy. Consequently, the detection of plant diseases is an essential research topic. In [28] an image-processing-based software solution is implemented for automatic detection and classification of plant diseases. In the first step of this study a color space transformation structure is created. In the second phase, k-means is used to segment the

images at hand. Furthermore, it was applied to partition the leaf image into four clusters in which one or more clusters contain the disease in case when the leaf is infected by more than one disease.

K-means algorithm is working only on numerical data and prohibits its use for clustering categorical data. In [29] the authors present two algorithms which to extend the k-means algorithm towards categorical domains and domains with mixed numeric and categorical values. The k-modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects and update the modes based on a frequency-based method in the clustering process to minimize the clustering cost function. The k-prototypes algorithm through the definition of a combined dissimilarity measure, further integrates the k-means and k-modes algorithms to allow for clustering objects described by mixed numeric and categorical attributes.

Most of clustering algorithms produce exclusive clusters meaning that each sample can belong to one cluster only. However, some medical datasets have inherently overlapping information which could be best explained by overlapping clustering methods. Overlapping k-means (OKM) is an extension of the traditional k-means algorithm that allows one sample to belong to more than one cluster. However, OKM also suffers from sensitivity to the initial cluster centroids. In [30] the authors propose a hybrid method that combines k-harmonic means and overlapping k-means algorithms (KHM-OKM) to overcome this limitation. The results of this study have shown that the proposed hybrid method provides better results compared to the original OKM algorithm. The effectiveness of the systematic initialization of OKM algorithm is demonstrated by comparing the objective function values at the first iteration of the OKM algorithm.

In [31], the authors propose a new k-means type smooth subspace clustering algorithm named Time Series k-means (TSkmeans) for clustering time series data which extracts smooth subspaces hidden in the data set. Also in this paper a new objective function is proposed to guide the clustering of time series data and the development of novel updating rules for iterative cluster searching with respect to smooth sunspaces.

Aspect-phrase grouping is an important task for aspect finding in sentiment analysis. [32] presented a flexible-constrained k-means algorithm to cluster aspect-phrases by using a user-specified threshold as lower bound on how well the given constraint must be satisfied. In computer vision field there has been increasing interest in learning hashing codes whose Hamming distance approximates the data similarity.

In [33] novel Affinity-Preserving K-means clustering algorithm is presented which simultaneously performs k-means clustering and learns the binary indices of the quantized cells.

The authors in [34] propose a new kind of k'-means algorithms for clustering analysis with three frequencies sensitive (data) discrepancy metrics in the cases that the exact number of clusters in a dataset is not pre-known. These algorithms can locate the centers of k' actual clusters by k' converged seed-points, respectively, with the extra k-k' seed points corresponding to empty clusters, namely containing no winning points in the competition according to the underlying frequency sensitive discrepancy metrics. Particularly these new k'means algorithms keep a simple learning rule, but have a rewarding and penalizing mechanism being similar to that of the rival penalized competitive learning algorithm.

Besides k-means, Self-organizing map (SOM) is also extensively used for the organization of the data. SOM is an unsupervised learning method that relates similar input vectors to the same region of a map of neurons. In essence, SOM (35) is a neural network that maps signal from a high-dimensional space to a one- or two-dimensional discrete lattice (M) of neuron units. Each neuron stores a weight. The map preserves topological relationships between inputs in a way that neighboring inputs in the input space are mapped to neighboring neurons in the map space. SOM mimics the clustering behavior observed in biological neural networks by grouping units that respond to similar stimuli together. Nerve cells, neurons, in the cortex of the brain seem to be clustered by their function. For example, brain cells responsible for vision form the visual cortex and those responsible for hearing form the auditory cortex.

In [36] a SOM is used to identify clusters in a large heterogeneous breast cancer database based on mammographic findings and patient age. The algorithm is used as a benchmark for model selection and to predict biopsy outcome. Disease infestation causes stress to the plant.

In [37] the research aims to detect stress and discriminate the type of stress from nutrient deficiency stress in field conditions using spectral reflectance information. After SOM is labeled, clusters of spectral features are identified which reflect the relationships between in the input environment. These SOM neurons are then able to estimate the stress status of an example spectrum presented to the SOM by calculating the Euclidean distance.

Another application of SOM concerns multi-disease diagnosis. In [38] the tomato disease features are extracted and a mapping relationship between the diseases and the features is created. Also the inaccurate clustering of traditional SOM algorithm has addressed with a two layers SOM models. Finally, SOM has been applied in socio economic studies.

Per capita ecological footprint (EF) is one of the most widely recognized measured of environmental sustainability. It seeks to quantify the Earth's biological capacity required to support human activity. In [39], SOM is used to model and cluster the EF of 140 nations. This study shows that SOM models are capable of improving clustering quality while extracting valuable information from multidimensional environmental data.

In association with self-organization, Mean Shift (MSH) is a powerful nonparametric technique that does not require knowledge of the number of clusters and does not constrain the shape of the clusters. MSH is based on the data density in the feature space in order to reveal areas of highest data concentration, which correspond to the modes of existing clusters. However, its performance suffers when the original metric fails to capture the underlying cluster structure. In [40] a semi-supervised framework for kernel MSH clustering is proposed and uses only pairwise constraints to guide the clustering procedure. The points are first mapped to a high-dimensional kernel space where the constraints are imposed by a linear transformation of the mapped points.

In [41] a mean shift-based clustering algorithm is proposed with three classes of Gaussian, Caushy and generalized Epanechnikov. Moreover, the proposed method aims at solving the bandwidth selection problems. In another article [42] the authors revisit Gaussian blurring mean-shift a procedure that iteratively sharpens a dataset by moving each data point according to the Gaussian mean-shift algorithm (GMS). A criterion is given in order to stop the procedure as soon

as clustering structure has been revealed, which produces image segmentations as good as those of GMS but faster.

The time complexity of the adaptive mean shift is related to the dimension of data and the number of iterations. In [43] an approximate neighborhood queries method is presented for the computation of high dimensional data in which, the locality-sensitive hashing is used to reduce the computational complexity of the adaptive mean shift algorithm. The data-driven bandwidth selection for multivariate data is used in mean shift procedure, and an adaptive MSH based on estimation algorithm is proposed.

## 1.4 Research on Cervical Cancer Diagnosis

Cervical cancer constitutes one of the most frequent types of cancer expressed in women worldwide, especially in woman under 40 years old [1]. It can be efficiently treated and cured when it is diagnosed in the first stages. In current clinical practice, the diagnosis of cervical cancer is mainly done through cervical screening followed by a necessary biopsy, but this method is labor consuming and expensive and can only detect superficial lesions around the external cervical orifice. In this way, there exists a wide scientific interest in prognosis, early diagnosis and treatment of precancerous lesions.

Colposcopy constitutes a conventional technique of cervix examination according to which a special magnifying device is utilized for the examination of the vulva, vagina and cervix areas and the identification of suspicious lesions, which are then biopsied and evaluated. It has proved effective in detecting malignancies around the external cervical aperture, however it may fail in recognizing all cancer lesions and is considered both time and labor expensive. Despite the achievements of colposcopy, cytology and medicine, via the introduction of new vaccines and vaccination policies that have resulted in reduced rates of cervical cancer morbidity and mortality, many lesions still remain undetected or overestimated leading to patients' health risk or their prompt to unnecessary biopsies respectively. Thus, reliable, cost effective and accurate tissue screening and testing methods must be utilized so as to catch more cases of the disease early, when it is most treatable. Pap smear test, optics, spectroscopy and high-resolution imaging methods are among the key directions for efficient cervical cancer screening [2, 3].Recent advances in medical research focus on the development of screening tests that are capable of overcoming the limitations of conventional cytology, with constitutes the gold-standard methodology for detection of cervical cancer in developed countries but is not easily applicable in regions where funding, techno structure and resources are limited [4].

Tissue evaluations take place considering the alteration of morphological and biochemical properties of the cervical sections and cells, indicating a malignancy evolution. Recently, a new technique originated from photoacoustic imaging (PAI), which is already being tested for the detection of skin and breast cancer [5]. The basic idea behind this approach lies on the irradiation of biological tissue via short laser pulses, which causes the conversion of the energy absorbed by it into heat and the generation of ultrasonic waves via the thermal

expansion within tissue. These generated signals are captured by an ultrasonic sensor to produce the corresponding images. The authors emphasized on the capability of photoacoustic imaging to distinguish cancerous from normal tissue and potentially evaluate the stage of the cancer, penetrating in higher depth and detecting lesions in the cervical canal, a region that conventional methods fail to efficiency screen. Although preliminary results confirmed this potential, further statistical validation on a larger testing dataset should be performed. Towards the direction of emerging approaches to cervical cancer screening, the authors in [6] proposed a novel technique for electrical characterization of cervical tissue, studying the bioelectrical properties of cells, which can reveal information on both their morphological and physiological characteristics.

When a cell is subjected to an electric field, it produces resistance to the current flow and reveals its bioimpedance properties, which vary under different applied frequency. The frequency response of the electrical bioimpedance of the tissue depends on its physiological and physiochemical status and is different from subject to subject, which forms a highly sensitive, spatiotemporal monitoring parameter for automated analysis of cellular behavior in vitro. This cellular property can be recorded utilizing cytosensors [7], which enable the conversion of cellular responses into a measurable electrical signal and the classification of cells as normal or abnormal ones in cancer screening via the electrical characterization of cervical exfoliative cytological samples. Although, this approach examines alternative features for tissue evaluation and succeeds in minimizing false negativity in the cervical cancer screening with Pap smear test, it is time consuming, since tissue sections must be collected and properly prepared for analysis and cannot reveal any information on the different cancer grades and evaluation stages of the disease.

Cervical cancer is expressed when abnormal cells on the cervix, the lower part of the uterus that opens into vagina, grow up in a rampant way. This kind of cancer can be treated successfully when detected in early stages, especially since screening tests and a vaccine to prevent the human papilloma virus (HPV) [8], the main cause of cervical cancer, are readily available. Cervical intraepithelial neoplasia (CIN) is believed that precedes invasive cervical cancer, which, when found early is highly treatable and associated with long survival and good quality of life. Chemical substances, as known as optimal biomarkers, are often used in order to increase the confidence of clinicians in cancer diagnosis and staging [9]. Imaging techniques are limited by the inherently weak optimal signals if endogenous chromospheres and fluorophores are used and also by the subtle spectral differences of normal and diseased biological samples. In the case of cervical cancer, topical application of acetic acid (AA) solution 3-5% is routinely used as a contrast agent for more than 70 years in order to highlight the abnormal areas [10]. The agent-tissue interaction generates an optical signal, which is perceived as transient tissue whitening. Clinical evidence supports that the degree and duration of the latter is associated with the lesion's grade, with the phenomenon known as acetowhitening (AW) effect. The method dictates the application of acetic acid to the cervix exterior for visualizing the biochemical reaction on it. The phenomenon can be observed under incandescent light, without magnification, producing a low intensity chemiluminescent, which allows the medical expert to have a subjective interpretation of the epithelial condition in vivo.

Optical technologies can improve the accuracy and availability of cervical cancer screening, allowing for both qualitative and quantitative analysis. Cancer screening involves the procedures of testing people, in most cases healthy ones, for signs that could reveal the early evolution of the disease and constitutes a means of cancer prevention. Detecting preliminary alterations in the neck of the womb could lead to early diagnosis and efficient termination of the malignancy evolution. It is desirable that this test is performed in a minimally invasive, easy, cost-effective and efficient way. Ferris t al. studied multimodal hyperspctral imaging for the noninvasive diagnosis of cervical neoplasia [11], reporting results of high sensitivity and specificity, while Huch et al. [12] evaluated the performance of optical detection of high=grade squamous intraepithelial lesion (HGSIL), a category of cervical dysplasia, using fluorescence and reflectance spectroscopy. Magnetic resonance imaging has also been reported in cervical cancer staging [13], providing extremely accurate and valuable findings at advanced stages of the disease and excellent imaging resolution for the different densities of pelvic structures.

In 2001, Balas [14] developed a novel multispectral imaging system, capable of performing time-resolved spectroscopy, for the *in vivo* early detection, quantitative staging and mapping of cervical cancer. This technique is based on measuring the modifications of the light scattering properties of the cervix, observed in cases of cervical neoplasia, after applying acetic acid solution to the examined tissue section. The processing and analysis of the optically enhanced output images revealed the increased sensitivity to detect incipient lesions, the priceless capability to extract additional, specific information regarding the evolution of the disease and the ability to discriminate neoplasias of different grade. A fulfill description on the progress and the benefits of biomedical optical imaging can be found in [15].

# References


[1] Irene M. Orfanoudaki, Dimitra Kappou, Stavros Sifakis, “Recent advances in optical imaging for cervical cancer detection”, Archives of Gynecology and Obstetrics, Vol. 284, Issue 5, pp 1197-1208, November 2011.

[2] Qiongshui Wu, Libo Zeng, Hengyu Ke, Hong Zheng, Xijian Gao, Diancheng Wang, "A multispectral imaging analysis system for early detection of cervical cancer", Proc. SPIE 5745, Medical Imaging 2005: Physics of Medical Imaging, 801, August 2005.

[3] Nadhi Thekkek and Rebecca Richards-Kortum, “Optical imaging for cervical cancer detection: solutions for a continuing global problem”, Macmillan Publishers Limited, Nature Reviews: Cancer, Vol 8, September 2008.

[4] Krishnakumar Duraisamy, K.S. Jaganathan and Jagathesh Chandra Bose, " Methods of Detecting Cervical Cancer", Advances in Biological Research, Vol. 5, No. 4, pp. 226-232, 2011.

[5] Kuan Peng, Ling He, Bo Wang, and Jiaying Xiao, "Detection of cervical cancer based on photoacoustic imaging—the in-vitro results", Biomedical Optics Express, Vol. 6, Issue 1, pp. 135-143, 2015.

[6] Lopamudra Das, Soumen Das and Jyotirmoy Chatterjee, " Electrical Bioimpedance Analysis: A New Method in Cervical Cancer Screening ", Journal of Medical Engineering, Volume 2015, Article ID 63607.

[7] B. Blad and B. Baldetorp, “Impedance spectra of tumour tissue in comparison with normal tissue: a possible clinical application for electrical impedance tomography,” Physiological Measurement, Vol. 17, No. 4, pp.A105–A115, 1996.

[8] Oliver Chukwujekwu Ezechi, Per Olof Ostergren, Francisca Obiageri Nwaokorie, Innocent Achaya Otobo Ujah and Karen Odberg Pettersson, "**T**he burden, distribution and risk factors for cervical oncogenic human papilloma virus infection in HIV positive Nigerian women", Virology Journal, Vol 11, No 5, January 2014.

[9] G C Giakos et.al., "Stokes parameter imaging of multi-index of refraction biological phantoms utilizing optically active molecular contrast agents", Measurement and. Science Technology, Vol 20, No 10, 104003, 2009.

[10] C. Balas, G. Papoutsoglou, and A. Potirakis, "In Vivo Molecular Imaging of Cervical Neoplasia Using Acetic Acid as Biomarker", IEEE J. Sel. Top. Quantum Electron. Vol. 14, pp 29-42 , 2008.

[11] D. G. Ferris, R. A. Lawhead, E. D. Dickman et al, “Multimodal hyperspectral imaging for the noninvasive diagnosis of cervical neoplasia”, J. Low. Genit. Tract Dis. Vol 5, No 2 ,pp 65–72, 2001.

[12] W. K. Huh, R. M. Cestero, F. A. Garcia et al, “Optical detection of high-grade cervical neoplasia in vivo: results of a 604 patient study,” Am. J. Obstet. Gynecol. Vol 190, pp 1249–1257 ,2004.

[13] Claudia C. Camisão, Sylvia M.F. Brenna, Karen V.P. Lombardelli, Maria Célia R. Djahjah, Luiz Carlos Zeferino, "Magnetic resonance imaging in the staging.

[14] Costas Balas, "A Novel Optical Imaging Method For The Early Detection, Quantitative Grading, and Mapping of Cancerous and Precancerous Lesions of Cervix", IEEE Transactions on Biomedical Engineering, Vol 48, No. 1, January 2001.

[15] Costas Balas, " Review of biomedical optical imaging—a powerful, non-invasive, non-ionizing technology for improving in vivo diagnosis ", Meas. Sci. Technol. , Vol. 20, Issue 10, 104020, 2009.

[16] A.K Jain, "Data Clustering: 50 Years Beyond K-means", Pattern Recognition Letters, Vol 31, pp 651:660, 2010.

[17] Cheng, Y., “Mean shift, mode seeking, and clustering”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 17, pp 790–799, 1995.

[18] Comaniciu, D. & Meer, P. “Mean shift: A robust approach toward feature space analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24, pp 603–619, 2002.

[19] Paulo JG Lisboa, Terence A Etchells, Ian H Jarman, Simon J Chambers, "Finding reproducible cluster partitions for the k-means algorithm", BMC Bioinformatics , Vol. 14, Suppl 1:S8, 2013.

[20] I. Vourlaki, G. Livanos, M. Zervakis, C. Balas, G. Giakos, "*Spectral Data Self-organization Based on Bootstrapping and Clustering Approaches*", Accepted for presentation in IEEE Imaging Systems and Techniques (IST) Conference, Macau, China, 16-18 September, 2015.

[21] Ke-Hai Yuan and Kentaro Hayashi, "Bootstrap approach to inference and power analysis based on three test statistics for covariance structure model", British Journal of Mathematical and Statistical Psychology, Vol 56, No 93, 2003.

[22] A. Irpino and R. Verde, "Dynamic clustering of interval data using a wasserstein-based distance," Pattern Recognition Letters, Vol. 29, No. 11, pp. 1648–1658, 2008.

[23] M. Chavent and Y. Lechevallier, "Dynamical clustering of interval data: optimization of an adequacy criterion based on hausdorff distance", in *Classification, Clustering and Data Analysis*, Springer, pp. 53-60, 2002.

[24] Fenfei Guo, Deqiang Han , Chongzhao Han, " *k*-intervals: a new extension of the *k*-means algorithm", Tools with Artificial Intelligence (ICTAI), IEEE 26th International Conference , pp. 251-258, Limassol, 2014.

[25] Elif Derya Ubeyli, Erdgan Doglu, "Automatic Detection of Erythemato-Squamous Diseases Using k-means Clustering", Journal of Medical Systems 34:179-184, 2010.

[26] Heidi S. Philips, Samir Kharbanda, Ruihuan Chen, William F. Forrest, Robert H. Soriano, Thomas D. Wu, Anhan Misra, Janice M. Nigro, Howard Colman, Liliana Soroceanu, P. Mickey Williams, Zora Modrusan, Burt G. Feuerstien and Ken Aldape, "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of diseases progression, and resemble stages in neurogenesis", Cancer Cell 9, 157-173, March 2006.

[27] John D. Shaughnessy Jr, Fenghuang Zhan, Bart E. Burungton, Yongsheng Huang, Simona Colla, Inchiro Hanamura, James P. Stewart, Bob Kordsmeier, Christopher Rndolph, David R. Wiiliams, Yan Xiao, Hongwei Xu, Joshua Epstien, Elias Anaissie, Somashekar G. Krishna, Michele Cottler-Fox, Klaus Hollming, Abid Mohiuddin, Maurucui Pineda-Roman, Guido Tricto, Frits van Rhee, Jeffrey Sawyer, Yazan Alsayed, Ronald Walker, Maurizio Zangari, John Crowley and Bart Barlogie, "A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome", Blood, Vol 109, No 6, 15 March 2007.

[28] Dheeb Al Bashish, Malik Braik and Sulieman Bani-Ahmad," Detection and Classification of Leaf Diseases using K-means based Segmentation and Neural-networks-based Classification", Information Technology Journal 10 (2):267-275, 2011.

[29] Zhexue Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery 2, 283-304, 1998.

[30] Sina Khanmohammadi, Naiier Adibeig, Samaneh Shanehbandy, "An improved overlapping k-means clustering method for medical applications" Expert Systems with Applications 67, 12-18, 2017

[31] Xiaohui Huang, Yumming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, Shaokai Wang, "Time series k-means: A new k-means type smooth subspace clustering for time series data", Information Sciences 367-368, 1-3, 2016.

[32] Shufeng Xiong, Donghong Ji, "Exploiting flexible-constrained k-means clustering with word embedding for aspect-phrase grouping", Information Sciences 367-368, 689-699, 2016.

[33] Kaiming He, Fang Wen, Jian Sun, "K-means Hashing: An Affinity-Preserving Quantization Method for Learning Binary Compact Codes", Conference Vision and Pattern Recognition, 2013 IEEE Conference.

[34] Chonglun Fang, Wei Jin, Jinwen Ma, "k'- Means algorithms for clustering analysis with frequency sensitive discrepancy metrics", Patterns Recognition Letters, Vol 34, Issue 5, Pages 580-86, April 2013.

[35] Kohonen, T, "Self-organizing maps". Berlin, Germany: Springer-Verlag, p. 501, 2001.

[36] Mia K. Markey, Joseph Y. Lo, Georgia D. Tourassi, Carey E. Floyd Jr, "Self-organizing map for cluster analysis of a breast cancer database", Artificial Intelligence in Medicine 27,113-127, 2003.

[37] D. Moshou, C. Bravo, S. Wahlen, J. West, “Simultaneous identification of plant stresses and diseases in arable crops using proximal optical sensing and self-organizing maps”, Precision Agric, 7:149-164, 2006.

[38] Ke Zhang, Yi Chai, Simon X. Yang, “Self-organizing feature map for cluster analysis in multi-disease diagnosis”, Expert Systems with Applications 37, 6359-6367, 2010.

[39] Mohamed M. Mostafa, “Clustering the ecological footprint of nations using Kohonen’s self-organizing maps”, Expert Systems with Application 37, 2747-2755, 2010.

[40] Saket Anand, Sushil Mittal, Oncel Tuzel, Peter Meer, “Semi-Supervised Kernel Mean Shift Clustering”, IEEE Transactions on pattern analysis and machine intelligence, Vol 36, No. 6, June 2014.

[41] Kuo-Lung Wu, Miin-Shen Yang, “Mean shift-based clustering”, Pattern Recognition 40, 3035-3052, 2007.

[42] Miguel A. Carreira –Perpinan, “Fast Nonparametric Clustering with Gaussian Mean-Shift, 23rd International conference on Machine learning, 2006.

[43] Xinhong Zhang, Yanbin Cui, Duoy Li, Xianxing Liu, Fan Zhang,” An adaptive mean shift clustering algorithm based on locality-sensitive hashing” Optik, 123, 1891-1894, 2012.

# 2. ALGORITHMIC FRAMEWORK FOR ORGANIZATION OF LARGE DATA SETS

## 2.1 Machine Learning Fundamentals

Alan Turing's proposal in his paper "Computing Machinery and Intelligence" that the question "Can machines think?" be replaced with question "Can machine do what we (as thinking entities) can do?"

Machine learning is a subfield of computer science and therefore, of artificial intelligence. In this way machine learning explores the study and construction of algorithms that can learn from and make predictions on data. In a heaven of applications and fields, re-emerging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, enhance computational processing that is cheaper and more powerful and affordable data storage.

Two of the most widely adopted machine learning schemes are supervised learning and unsupervised learning. Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. Unsupervised learning, which is the tool of our study, is used against data that are not accompanied by historical labels. The goal is to explore the data and find some structure within. A third method is reinforcement learning, where a computer interacts with a dynamic environment in which it pursues a certain goal, without a teacher explicitly it whether it has come close to its goal.

This thesis is primarily concerned with the organization of data without prior information. Its bases, thus, are on the foundations of unsupervised learning using clustering techniques and the focus is on how we can use data more efficiently within clustering in order to reveal hidden organizational principles that guide the formation of such data.

## 2.2 Data Clustering Fundamentals

Data clustering is, nowadays, a well-known process of partitioning or grouping a given set of patterns into classes of similar objects. Clustering or cluster analysis, as it is called, is an unsupervised method in contrast to classification that is a supervised method. It is a main task of exploratory data mining and a technique for statistical data analysis. Furthermore, cluster analysis has many applications and it is used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics and data compression. Clustering can also help market analysts discover distinct groups in their customer base. Also, they can characterize their customer groups based the purchasing patterns. Furthermore, clustering can be used in the field of biology to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures of populations.

The main goal of clustering analysis is to reveal the natural groupings of a set of patterns, points or objects. The objects within a group will be similar (or related) to one another and different (or unrelated) to objects in other groups. The greater the similarity within a class and the difference among classes, the better or more distinct the clustering result. A general definition of clustering can be stated as follows: Given a database of n objects, find K groups based on a measure similarity, such that the similarities between objects in the same group are high, while the similarities between objects in different groups are low [1]. The main clustering steps are the following [2,3]:

1. Pattern representation (optionally including feature extraction or selection),
2. Definition of a pattern proximity measure appropriate to the data domain,
3. Grouping,
4. Data extraction (it is not necessary),
5. Cluster validity (it is not necessary)

*Pattern representation* relates with the number of classes, the number of available patterns, which are available to the clustering algorithm.

*Feature selection* is the process of recognition that is the most effective subset of the original features to use in clustering. It also contributes to smaller training times and reduces variation. The general idea behind the feature selection technique is that the data contains many features that can often be redundant and so can be removed without losing information.

*Feature extraction* *creates* new features of operations on the original features and the feature selection returns a subset of them.

*Pattern proximity* depends on the minimization of a distance criterion. The most useful distance measure is the Euclidean distance.

Mathematically, the definition of Euclidean distance between two n-dimensional vectors $x = (x_1, ...x_n)$ and $y = (y_1..., y_n)$ is [5]:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad 2.1$$

*The grouping step* has many different implementations and depends on the type of clustering. The output clustering can follow a partitioning method, where the algorithms identify the partition that optimizes a clustering criterion. The most famous algorithm in this category is the k-means that we develop below. Alternatively, fuzzy clustering can be used, in which each object has a degree of membership in each of the producer cluster. A third choice is the hierarchical methods, that create a hierarchical decomposition of the given data objects.

*Data abstraction* is the reduction of a particular body of data to a simplified representation of the whole. In general, it is the process of taking away or removing features from something

in order to reduce it to a set of essential characteristics. In clustering terms, a typical data abstraction is a compact description of each cluster, the most times in terms of cluster prototypes or representative patterns such as the centroid [2, 4].

Cluster validity is an essential step following the clustering process, which assesses the quality of the clustering process. Cluster validity expresses what makes a clustering result good or better from another clustering. Several validity approaches have been developed, some of the most important ones being the following:

- *Dunn Index*
- *Davies Bouldin Index*
- *Silhouette criterion*

*Figure 1* depicts a sequence of training process, including a feedback path where the grouping process results could affect the feature extraction and similarity measure.

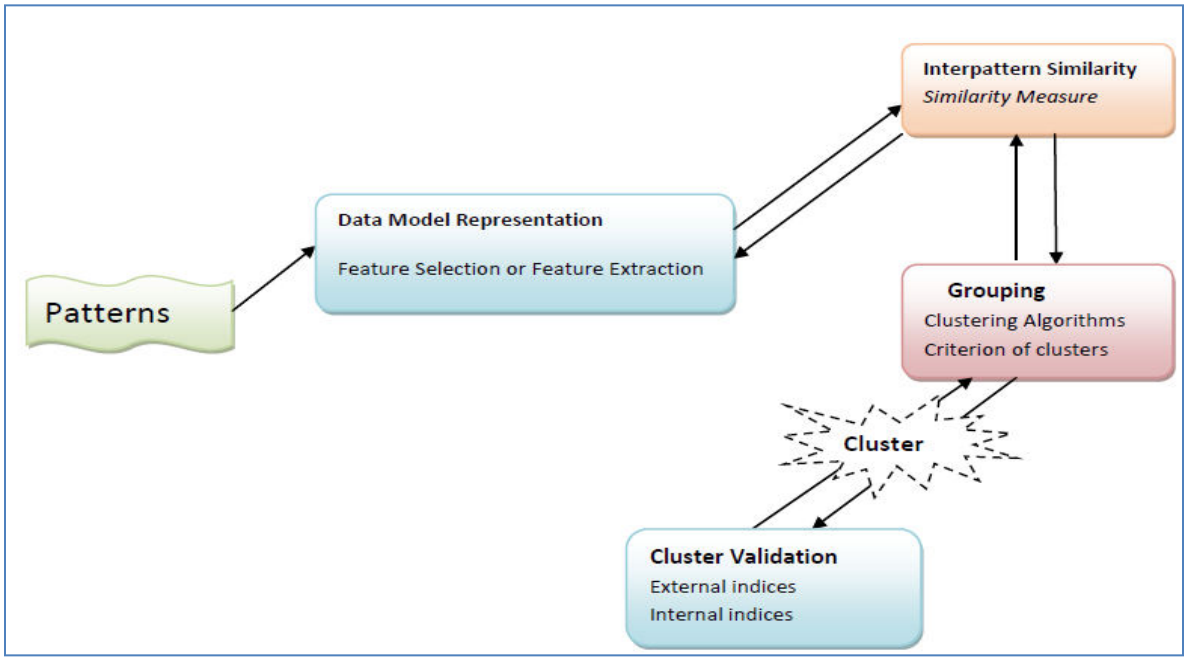


**Figure 1**. Clustering process

## 2.3 Clustering Techniques

How successful a clustering process can be depends on the choice of features, the choice of similarity measure and especially the choice of suitable data organization. Different approaches to clustering data have been developed during the years with three basic categories of algorithms dominating to the rests:

1. Partitioning Relocation Clustering

2. Hierarchical Clustering
3. Fuzzy clustering

### 2.3.1 Partitioning Relocation Clustering

Partitional clustering algorithms obtain a single partition of the data instead of a clustering structure. They consider a determined number of groups (clusters) and assign the data into these groups. Finally, they aim to optimize the result. They use a $n \times d$ matrix, where $n$ objects are grouped in a d-dimensional feature space or a $n \times n$ similarity matrix. All data points are assigned to the closest group. Following that, the position of each point is redefined by an iterative mode until the similarity criterion converges. In each group a value is assigned which is gradually minimized [2, 5]. This value is represented by the within-cluster sum of squares (sum of distance functions of each point in the cluster to the center respectively). The biggest problem of a partitional algorithm is the choice of the number of desired output clusters. They produce clusters by optimizing a criterion function defined either locally (on subset of the patterns) or globally (defined over all of the patterns). The taxonomy of clustering approaches is shown in Figure 2.

The most widely employed criterion function in partitional clustering techniques is the squared error criterion, which tends to work well with isolated and compact clusters. The squared error of a clustering of a pattern set and K clusters is:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j \right\|^2 , \qquad 2.2$$

where $x_i^{(j)}$ is the $i^{th}$ pattern belonging to the $j^{th}$ cluster and $c_j$ is the centroid of the $j^{th}$ cluster.

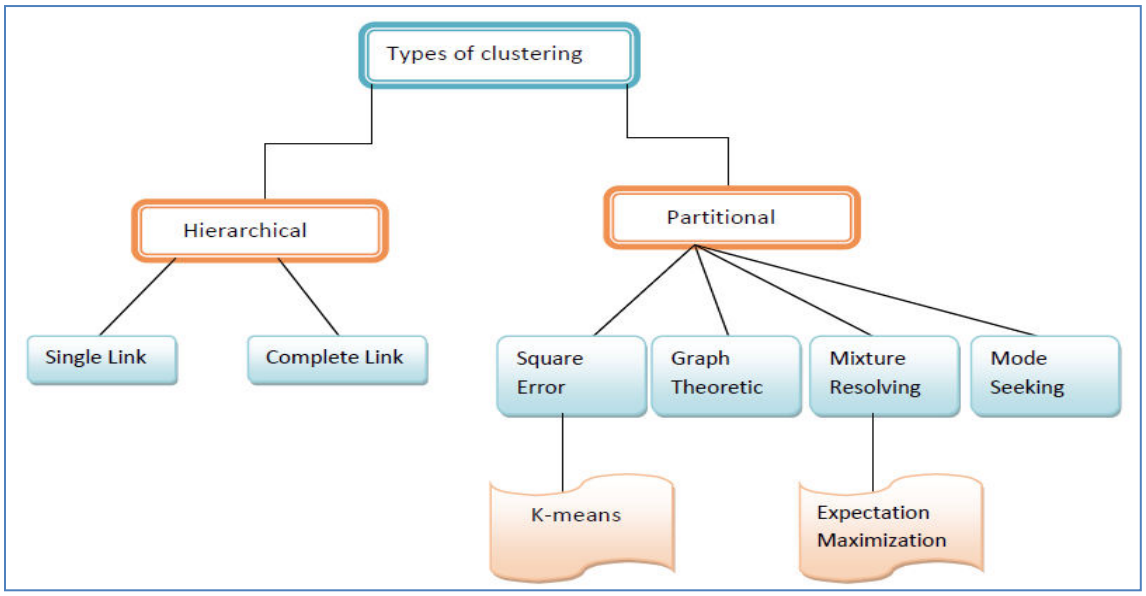


**Figure 2.** Clustering techniques

The k-means is the most common and popular algorithm [McQueen 1967]. K-means often employs a squared error criterion. It starts with a random initialization and keeps reassigning each point to clusters based on the similarity between the points and the cluster centers until the distance criterion converges and there is no reassignment of any point from one cluster to another. The k-means algorithm is famous because is easy to implement and its time complexity is $O(n)$, where $n$ is the number of patterns [2]. A huge problem with this algorithm is that is sensitive to the initialization (this problem will be analyzed in a following chapter) and may converge to a local minimum of the criterion function value, if the initial partition is not properly chosen. The basic steps of squared error clustering method are described below:

- Chose randomly an initial partition of the patterns, having a standard number of clusters and centers.
- Move to assignment, produce for every point to the closest cluster center and compute the new cluster center. After, proceed to repeat this step until convergence is achieved and the cluster membership stop changing.
- Adjust the number of clusters by, merging and splitting existing clusters or by removing small or outlier clusters.

### 2.3.2 Hierarchical algorithms

Hierarchical algorithms aim to create a hierarchy within the data. They create a dendogram indicating the number of groups and size allowed. The methods widely utilized belong to the family of sequential fission (agglomeration) or fusion (division) methods that contrast the hierarchy level-by-level, from bottom to top (agglomerative clustering) or from top to bottom (divisive clustering) [8, 9]**.**

*Agglomerative clustering* results can be represented in a tree form. Initially it assumes that each data point is itself a cluster and it contains only itself. Then it merges the most similar pair groups sequentially, to create a hierarchy of groups. The most popular agglomerative algorithms are the single link, complete link, average link, Ward's (Incremental Sum of Squares) [10]. The first three of the methods can be applied to any dissimilarity /similarity data while the Ward's method is developed for the entity –to-variable data (using between-centroids distances).

*Divisive clustering* starting with all data points in one cluster and successively dividing each cluster into smaller cluster.

Comparing partitional algorithms with hierarchical ones, the partitional algorithms find all clusters in the same time and do not impose a hierarchical structure. The input data is a similarity matrix $n \times n$, where $n$ is the number of objects to be grouped.

### 2.3.3 Fuzzy clustering

Clustering can be classified as Soft clustering (Overlapping Clustering) and Hard Clustering (or Exclusive Clustering). In hard clustering each object has two options, to belong or not in one

cluster. Opposite, in the case of soft clustering the objects may belong to two or more clusters with different degrees of membership. In this option, data will be associated to an appropriate membership value. This means that each cluster contains memberships and each of them is characterized by a degree value between 0 and 1. The fuzzy criterion function, e.g., a weighted squared error criterion function can possible is [2]:

$$Q = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ij} \left\| x_i - c_k \right\|^2 , \qquad 2.3$$

where $c_k = \sum_{i=1}^{N} u_{ik} x_i$ is the $k^{th}$ fuzzy cluster center.

The most famous algorithm of this category is the Fuzzy C Means (FCM). FCM is a data clustering technique where each object belongs to a cluster to some degree that is specified by a membership grade. This technique was introduced by Jim Bezdek in 1981 [1, 6]. A basic difference between FCM and K-means is that FCM is taking more time for computation than that of K-means. The time complexity of K-mean algorithm is $O(ncdi)$ and time complexity of FCM is $O(ndc^2 i)$ [7].

## 2.4 Thesis explored algorithms

In this study two algorithms are chosen:

1. k-means
2. Mean shift.

K-means is one of the most widely used algorithms in machine learning and Mean shift (MSH) is a non-parametric technique based on an empirical probability density function. K-means is used to produce many centers of our data, while Mean shift to self-organize all these and evaluate the optimal number of clusters. Before we analyze the main goal of our study and methodology, a basic description of these two algorithms follows.

### 2.4.1 K-means an efficient distance based algorithm

K-means was first published in 1955 [1]. Since it was introduced, is still remains one of the most popular and simple clustering algorithms. Ease of implementation, simplicity, efficiency and empirical success are the main reasons for its popularity.

*What is its goal?*

K-means is a method that describes the best possible partitioning of a data set containing $k$ number of clusters. The method is defined by its objective function which aims to minimize the sum of all squared distances within a cluster, for all clusters. The objective function is defined as:

$$\arg\min_{S} \sum_{i=1}^{k} \left( \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \right) \qquad 2.4$$

where $x_j$ is a data point in the dataset, $S_i$ is a cluster (set of data points) and $\mu_i$ is the cluster mean (the center point of cluster $S_i$).

Minimizing this objective function is known to be an NP-hard problem [11]. K-means can only converge to a local minimum, even though recent studies have shown with a large probability, K-means could converge to the global optimum when clusters are well separated [12].

One aspect of k-means that makes it different from many other clustering methods is that the number of clusters is fixed when clustering occurs. This can be considered both as a weakness and strength. One positive consequence of a fixed number of clusters is that the k-means method does not introduce new cluster in case of an anomaly data point, instead it sorts the anomaly data point to its closest cluster. The drawback of using a fixed number of clusters is that it might not be clear how many clusters a dataset might contain. Using an unsuitable $k$ may cause the k-means method produce poor results, possibly to the point of becoming unusable. As with any clustering method k-means is not suitable for all types of data. Even the case individual clusters have suitable properties for k-means clustering the density and position of the cluster can affect the result.

*Algorithmic Methodology*

The algorithmic method consists of two separate sections. The first phase is to take each point that belongs to the given data set and associate it to the nearest centroid. The most widely useful distance is the Euclidean distance that is generally considered to determine the distance between data points and the centroids. When all the points are assigned to clusters, the first step is completed. After that, we need to recalculate the new centroids and because of that, the first phase is repeated until new centroids do not change their position. This signifies the convergence criterion for clustering. The basic algorithm steps are following:

| Algorithm 1: The k-means clustering algorithm |
|---|
| *Input:*<br><br>$X = \{x_1, x_2, .., x_n\}$ // set of $n$ data items.<br>$k$ // Number of desired clusters<br><br>*Output:*<br><br>A set of $k$ clusters. |

*Steps:*
Select randomly $k$ instances as initial cluster centers.

*Repeat*
Assignment step: Assign each observation $d_i$ to the cluster whose mean yields the least within cluster sum of squares. Since the sum of squares is the Euclidean distance, this is intuitively the nearest mean.

*Update step*: Calculate new mean for each cluster;

*Until convergence criteria is met*

*K-means parameters*

Three user- specified parameters are required for k-means algorithm:

1. Number of clusters $k$
2. Cluster initialization
3. Distance metric

The most critical choice is $k$. While no perfect mathematical criterion exists, a number of heuristics (1,13) are available for choosing $k$. K-means is run independently for different values of $k$ and the partition that appears to be the most meaningful to the domain expert is selected. K-means although relatively fast, adaptive and effective, has the drawback that different location of centers can cause different results, thus the algorithm suffers from initialization. Correcting these standard errors is a challenging issue in statistical analysis. K-means is typically used with the Euclidean distance computing the distance between points and cluster centers. As, a result K-means finds spherical or ball-shaped clusters in data. For high-dimensional data, the Euclidean distance is less meaningful in such a spherical space than the Cosine similarity or Pearson correlation, which is used in the spherical k-means. With these three parameters and effects we will deal in the sequel.

*K-means Properties*

After a basic description of k-means attribute, we can separate its characteristics at advantages and disadvantages [14, 15, and 16]:

*Advantages*

- ✓ *Fast, robust, easy to implement.*

- ✓ *Relatively efficient:* $O(tknd)$, *where* $n$ *is objects,* $k$ *is clusters, d is dimension of each object and* $t$ *is the number of iterations. Normally, k, t, d<<n. k-means is linear in all relevant factors (iterations, number of clusters, number of documents and dimensionality of the space).*
- ✓ *Gives best results when data are compact and well separated from each other (Figure 3)*

*Disadvantages*

- *The learning algorithm requires a priori knowledge of the number of cluster centers.*
- *Sensitive to the initialization, provides the local optima of the squared error function.*
- *The use of Exclusive Assignment: if there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.*
- *Algorithm fails from non-linear data set (Figure 3).*

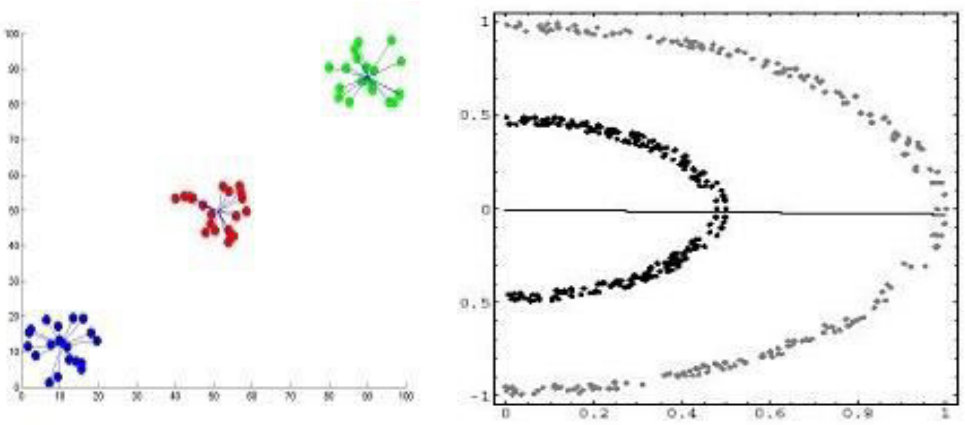

**Figure 3:** First image (left) illustrates the results of k-means for k=3 and well separated clusters, while the second image (right) depicts the non-linear data set [16]

### 2.4.2 MEAN SHIFT ALGORITHM

MSH is a powerful non parametric iterative algorithm that can be used for lot of purposes like finding modes and clustering and does not require prior knowledge of the number of clusters. Mean shift was introduced by Fukunaga and Hosteter in 1975 [17], was later adapted by Cheng [18] for the purpose of image analysis and has been extended to be applicable in other fields like Computer Science.

*MSH idea*

The main idea behind MSH is to model points in the $d-$dimensional feature space as an empirical probability density function, where dense regions in the feature space correspond to the local maxima or modes of the underlying distribution. If the input is a set of points, then MSH considers them as sample from the underlying probability density function. For each data

point in the feature space, MSH associates it with the nearby peak of the dataset's probability density function. Furthermore, for each data point, MSH creates a window around it and then calculates the mean of the data point, that include on this window. Then it shifts the window to the defined mean of the data point. The algorithm repeats this procedure until the window stops moving and the algorithm converges. Data points that associate with the same stationary point are considered members of the same cluster [19].

*MSH methodology*

The basic framework of MSH operation is following:

- *Fix a window around each data point.*
- *Compute the mean of data within the window*
- *Shift the window to the mean and repeat till* convergence.

Assume that we have a set of points in two dimensional spaces, and we draw a circle somewhere in this space and at least one of the points is inside this circle. This circle is called ''kernel'' in the Mean shift language. The radius of this circle is called ''bandwidth''. The bandwidth is essentially the only parameter of the Mean-shift method except the choice of the kernel.

*Mean shift principle*

Cheng in his paper [18] generalizes and analyzes the Mean shift algorithm as a mode-seeking process on a surface constructed with a ''shadow'' kernel. For Gaussian kernels, Mean shift is a gradient mapping.

First we must define the definition of kernel. So that a kernel is a function that satisfies the following requirements [19]:

$$1.\int_{R^d} \varphi(x) = 1 \qquad 2.5$$

$$2.\varphi(x) \geq 0 \qquad 2.6$$

Given n data points, $x_i \in \mathbb{R}^d$ the multivariate kernel density estimate obtained with kernel $K(x)$ and window radius h (termed the bandwidth parameter) is,

$$\hat{f}_K = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad 2.7$$

For radially symmetric kernels, it suffices to define the profile of the kernel $k(x)$ satisfying

$$K(x) = c_{k,d} k\left(\|x\|^2\right) \tag{2.8}$$

where $c_{k,d}$ is a normalization constant with assures $K(x)$ integrates to 1. The modes of the density function are located at the zeros of the gradient function $\overline{\nabla} f(x) = 0$.

The gradient of the density estimator (1) is

$$\hat{\overline{\nabla}} f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x)\, g\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \tag{2.9}$$

$$= \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)\right] \left[\frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x\right]. \tag{2.10}$$

where $g(s) = -k'(s)$. The first term is proportional to the density estimate at $x$ computed with kernel $G(x) = c_{g,d}\, g\left(\|x\|^2\right)$ and the second term

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \tag{2.11}$$

is the mean shift. The mean shift vector always points toward the direction of the maximum increase in the density and is proportional to the density gradient estimate at point $x$ obtained with kernel $K$. The mean shift procedure for a given point $x_i$, obtained by successive:

1. Computation of the mean shift $m\left(x_i^t\right)$.
2. Translation of the window $x^{t+1} = x^t + m_h\left(x^t\right)$
3. Iteration of the steps 1 and 2 until convergence, i.e., $\overline{\nabla} f(x_i) = 0$.

The set of all locations that converge to the same mode defines the basin of attraction of that mode. The points which are on the same basin of attraction could belong to the same cluster [20]. The principle of MSH is illustrated in Figure 4 [21].

*MSH issues*

- MSH algorithm is time intensive. The time complexity of it is given by $O\left(Tn^2\right)$ where $T$ is the number of iterations and n is the number of data points in the data set.
- MSH depends on the value of bandwidth parameter h that is required. One way to calculate the optimal value of is using k-nearest neighbor. The convergence of the algorithm and the number of the clusters are influenced by the h.
- The variation of the h parameter affects to the clusters. A large h might be result in incorrect clustering and might merge separate clusters. On the other hand a smaller choice of value h might result in too many clusters.
- MSH is not so efficient in higher dimensions, because the number of local maxima is very high and it might converge to local optima soon.

### *MSH and clustering*

Mean shift's most important application is for clustering. That is because the algorithm does not require the knowledge of number of clusters or the shape of the clusters. As we notice above, Mean shift is a mode seeking algorithm that we can utilize to find clusters. The stationary points obtained via gradient ascent represent the modes of the density function. All points associated with the same stationary belong to the same cluster.

The main difference of that algorithm with K-means, is that the latter makes two wide assumptions, the number of the clusters and the requirement of the spherical shape of clusters. Also k-means is very sensitive to initializations. This means that a bad initialization can delay convergence or even lead to wrong clusters. From the other hand Mean shift is quite robust to initializations. K-means is faster from Mean shift, which is computationally expensive.

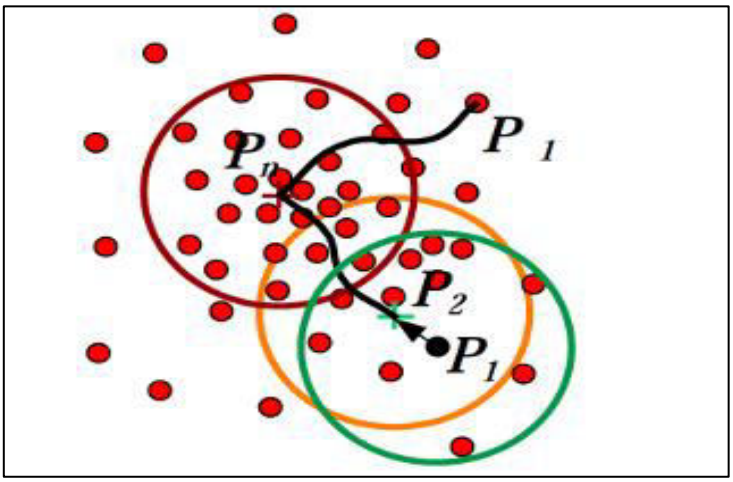


**Figure 4: Principle of mean shift algorithm [21].**

## 2.5 Distance Metrics

Since clustering is the grouping of similar objects, some measures that can determine whether two objects are similar or dissimilar are required. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. It is useful to denote the distance between two points $x_1$ and $x_2$ as: $d(x_1, x_2)$. A valid distance measure should be symmetric and obtains its minimum value (usually zero) in case of identical vectors. The distance measure is called a metric distance measure if it also satisfies the following properties [22]:

1. Triangle inequality $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k) \quad \forall x_i, x_j, x_k \in S.$
2. $d(x_i, x_j) = 0 \Rightarrow x_i = x_j \quad \forall x_i, x_j \in S$.

### 2.5.1 Euclidean distance

The most common distance which is used is the Euclidean distance. In general, if we have $p$ variables $X_1, X_2, ..., X_p$ measured on a sample of n subjects, the observed data for subject $i$ can be denoted by $x_{i1}, x_{i2}, ...x_{ip}$ and the observed data for subject $j$ by $x_{j1}, x_{j2}, ...x_{jp.}$ the Euclidean distance between these two subjects is the following:

$$d_{i,j} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \qquad 2.12$$

Euclidean distances are usually computed from raw data and not from standardized data. The main advantage about this distance is that it does not depend on the addition of new objects to the analysis, which may be outliers. Against this, Euclidean distance is affected from differences in scale among the dimensions from which the distances are computed.

In addition, if one variable has a much wider range than others then this variable will tend to dominate. For example, if one of the dimensions denotes a measured length in centimeters and we then convert it to millimeters the resulting Euclidean can be greatly affected. To get around this problem each variable can be standardized (converted to z-scores). In consequence the results of cluster analyses may by different, as it tends to reduce the variability (distance) between clusters. This happens because if a particular variable separates observations well, then it will have a large variance (as the between cluster variability will be high). If this variable is standardized, then the separation between clusters will become less. Generally, it is a good practice to transform the dimensions so that they have similar scales [22, 23].

### 2.5.2 Cosine Distance

An alternative concept to that of the distance is the similarity function $s\left(x_i, x_j\right)$ that compares the two vectors $x_i$ and $x_j$. This function should be symmetrical (like $s\left(x_i, x_j\right) = s\left(x_j, x_i\right)$) and have a large value when $x_i$ and $x_j$ are somehow similar and constitute the largest value for identical vectors. A similarity function where the target range is [0, 1] is called a dichotomous similarity function. Specifically, the cosine similarity between two vectors is a measure that calculates the cosine of the angle between them. This metric is a measure of orientation and not magnitude. Given a n m-by-n matrix X, which is treated as m (1-by-n) row vectors $x_1, x_2, ..., x_m$, the cosine similarity between the vector $x_s$ and $x_t$ is following [23, 24):

$$\text{Similarity (x, y)} = \frac{x_s x_t'}{\sqrt{\left(x_s x_s'\right)\left(x_t x_t'\right)}} \qquad 2.13$$

The cosine distance between two pints is one minus the cosine of the included angle between points (treated as vectors). This equation is following:

$$D = 1 - \frac{x_s x_t'}{\sqrt{\left(x_s x_s'\right)\left(x_t x_t'\right)}} \qquad 2.14$$

As mentioned before, cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Thus, it is a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1 and two vectors diametrically opposed have a similarity of -1, independently of their magnitude.

# References


1. A.K. Jain, “Data Clustering: 50 Years Beyond K-means”, Pattern Recognition Letters, Vol 31, pp 651:660, 2010.
2. A.K. Jain, M.N. Murty, P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol 31, No 3, September 1999.
3. A.K. Jain, R.C. Dubes, “Algorithms for Clustering Data”, Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ, 1988.
4. E. Diday, J. C. Simon, “Clustering Analysis”, Digital Pattern Recognition, K.S. Fu, Ed, Springer-Verlag, Secaucus, NJ, pp 47:94.
5. B. Mirkin, "Mathematical Classification and Clustering", Kluwer Academic Publishers Group, 1996
6. M. Sato, Y. Sato, L.C. Jain, “Fuzzy Clustering models and Applications”, Physical-Verlag, GmbH & Co, 1997.
7. Sumi Ghosh, Sanjay Jumar Dubey,” Comparative Analysis of K-Means and Fuzzy C-Means Algorithms”, International Journal of Advanced Computer Science and Applications, Vol 4, No 4, 2013.
8. Pavel Berkhin, “Survey of Clustering Data Mining Techniques”, Accrue Software, Inc.
9. G. Lance, W. Williams, “A general theory of classification sorting strategies”, Computer Journal, Vol 9, pp 373:386, 1967.
10. J.H. Ward, “Hierarchical grouping to optimize an objective function”, Journal Amer. Stat, Assoc., Vol 58, No 301, pp 235:244, 1963.
11. P. Drineas, A. Frieze, R. Kanna, S. Vempala, V. Vinay, “Clustering large graphs via the singular value decomposition”, Machine Learn, Vol 56, No 1:3, pp 9:33, 1999.
12. Melia Marina, “The uniqueness of a good optimum for k-means”, Prpc., 23rd, Internat, Conf, Machine Learning, pp 625:632, 2006
13. R. Tibshirani, G. Walther, T. Hastie, “Estimating the number of clusters in a data set via the gap statistic”, J. Roy Statist, Soc, B, pp 411:423, 2001.
14. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angel Y. Wu, “An efficient k-Means Clustering Algorithm: Analysis and Implementation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24, No 4, July 2002.
15. KeChen,“k-meansclustering” https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxkYXRhY2x1c3RlcmluZ2FsZ29yaXRobXN8Z3g6NDkxNDNmZGUxMzE5YzgyNg.
16. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.
17. K. Fukunaga, L. Hosteler, “The estimation of the gradient of a density function, with applications in pattern recognition”, IEEE Transactions on Information Theory, Vol 21, No 1, pp 32:40, 1975.
18. Y. Cheng, “Mean shift, mode seeking and clustering”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 17, No 8, pp 790-799, 1995.
19. Kostantinos G. Derpanis, “Mean shift Clustering”, August 15, 2005.
20. D. Comaniciu, V. Ramesh, P. Meer, “A Robust Approach Toward Feature Space Analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24, No 5, May 2002.
21. http://sociograph.blogspot.gr/2011/11/accessible-introduction-to-mean-shift.html.
22. Lior Rokach, Oded Maimon, “Data Mining and Knowledge Discovery Handbook”.
23. Dibya Jyoti Bora, Dr Anil Kumar Gupta, “Effect of Different Distance Measures on the Performance of K-means Algorithm: An Experimental Study in Matlab”, International Journal of Computer Science and Information Technologies, Vol 5, No 2, pp 2501:2506, 2014.
24. Jianfeng Li, D Brynn Hibbert, Stephen Fuller, Gary Vaughn, “A Comarative Study of Point-to-Point Algorithm for Matching Spectra”.

## 2.6 Summary of application in time series mining

Times series data constitute an important category of temporal data objects and often acquired from scientific and financial applications. A time series is a collection of observations made chronologically. The nature of time series data relates to large data size, high dimensionality and necessity to update continuously. Furthermore, time series, which is characterized by its numerical and continuous nature, is always considered as a whole instead of individual numeric field. The increasing use of time series has initiated a great deal of research and development attempts in the field of data mining. Such data mining research is categorized into representation and indexing, similarity measure, segmentation, visualization and mining [1]. Based on the time series representation, different mining tasks can be found in the literature and they can be classified into four fields: pattern discovery and classification, classification, rule discovery and summarization.

### 2.6.1 Representation and Indexing

A fundamental problem in the field of time series mining is how to appropriately represent the temporal dependence of data. One of the major aims for time series representation is to reduce the dimension (i.e the number of data point) of the original data. Sampling [2] is the simplest method for dimension reduction. In this method, a rate of m/n is used, where m is the length of a time series P and n is the dimension after dimensionality reduction. However, the sampling method has a negative effect in the shape of sampled time series, if the sampling rate is too low. Another method which is called Piecewise Aggregate Approximation (PAA) uses the segmented means to represent the time series [3]. An extended version of PAA is s proposed by [4], in which the length of each segment is not fixed but adapted to the shape of the series and is called Adaptive Piecewise Constant Approximation (APCA). Except of using the mean to represent each segment, other methods propose for instance, to use the segmented sum of variation (SSV) to represent each segment of the time series [5].

In order to reduce the dimension of time series data, another approach is to approximate a time series with straight lines. Two major categories are included. The first one is linear interpolation. A common method is using piecewise linear representation (PLR) [6]. It tends to closely align the endpoint of consecutive segments, giving the piecewise approximation with connected lines. PLR is a bottom-up algorithm. It begins with creating a fine approximation of the time series, so that m/2 segments are used to approximate the m length time series and iteratively merges the lowest cost pair of segments, until it meets the required number of segment. The second approach is linear regression, which represents the subsequences with the best fitting lines [7].

Furthermore, reducing the dimension by preserving the salient points is a promising method. These points are called as perceptually important points (PIP). The PIP identification

process was first introduced by [8] and used for pattern matching of technical (analysis) patterns in financial applications [9]. In [10] a lattice structure is proposed to represent the identified peaks and troughs (called control points) in the time series. A critical point model [11] and a high-level representation based on a sequence of critical points [12] are proposed for financial data analysis.

One the other hand, special points are introduced to restrict the error on PLR [13]. Key points are suggested to represent time series in [14] for anomaly detection. Another common family of time series representation approaches converts the numeric time series to symbolic form. That is, first discretizing the time series into segments, then converting each segment into a symbol [15]. Moreover, in [16], the authors propose to represent each segment by a codeword from a codebook of key-sequences. This work has extended to multi-resolution consideration [17].

Furthermore, subsequence clustering is a common method to generate the symbols [18]. A multiple abstraction level mining approach is proposed by [19], which is based on the symbolic form of the time series. The symbols in this paper are determined by clustering the features of each segment, such as regression coefficients, mean square error and higher order statistics based on the histogram of the regression residuals.

Most of the methods described so far are representing data in time domain directly. Representing time series in a transform domain is another large family of approaches. One of the most popular transformation techniques in time series mining is the discrete Fourier transforms (DFT), since first being proposed for use in this context by [20]. Moreover, in [21] is proposed to use likelihood statistics to test the hypothesis of difference between series instead of an Euclidean distance in the transformed domain.

Principal component analysis (PCA) is a popular multivariate technique used for developing multivariate statistical process monitoring methods [22]. In most related works, PCA is used to eliminate the less significant components or sensors and reduce the data representation only to the most significant ones and to plot the data in two dimensions.

Many of the representation schemes described above are incorporated with different indexing methods. A multi-level distance based index structure is proposed [23], which for indexing time series represented by PCA.

As consequence of the above approaches for a given index structure, the efficiency of indexing depends only on the precision of the approximation in the reduced dimensionality space. However, in choosing a dimensionality reduction technique, we cannot simply choose an arbitrary compression algorithm but we rather seek for a technique that produces an indexable representation. For example, many time series can be efficiently compressed by delta encoding, but this representation does not lend itself to indexing. In contrast, DFT and PCA all lend themselves naturally to indexing, with each Fourier coefficient or aggregate segment map onto one dimension of an index tree. Then, the processing is performed by computing the actual distance between sequences in the time domain and discarding any false matches.

### 2.6.2 Similarity measure

Similarity measure is of fundamental importance for a variety of time series analysis and data mining tasks. In time series data, which is characterized by its numerical and continuous nature, similarity is typically carried out in a user defined manner. To measure the similarity/dissimilarity between two-time series, the most popular approach is to evaluate the Euclidean distance on the transformed representation like the DFT coefficients [24]. Besides Euclidean-based distance measures, other distance measures can easily be found in the literature. One of the most popular and field-tested similarity measures is called the "time warping" distance measure. Based on the dynamic time warping (DTW) technique, the proposed method in [25] predefines some patterns to serve as templates for the purpose of pattern detection. Focusing on similar problems as DTW, the longest Common Subsequence model [26] is proposed. The Common Subsequence model is a variation of the edit distance and the basic idea is to match two sequences by allowing them to stretch, without rearranging the sequence of the elements, but allowing some elements to be unmatched. A parameter-light distance measure method based on Kolmogorov complexity theory is suggested in [27].

In subsequence matching where a query sequence and a longer time series are compared, the task is to find the subsequences in the longer time series that matches the sequence. The General Match method is proposed by [28], which reduces the window size effect by using large windows by the method in [29] and exploits point-filtering effect by Dual Match [30]. Detailed comparisons and experiments on the existing time series representation and similarity measure approaches can be found in [31, 32].

### 2.6.3 Segmentation

Time series segmentation can be considered either as a preprocessing steps for variety of data mining tasks or as trend analysis techniques. In [33], a simple discretization method is proposed. A fixed length window is used to segment a time series into subsequences and the time series is then represented by the primitive shape patterns that are formed. Exploiting attributes of on PCA, fuzzy clustering based segmentation based segmentation is proposed in [34]. The segmentation problem has also been considered from the perspective of finding cyclic periodicity for all of the segments. In [35], the data cube and the Apriori data mining techniques are used to mine segment-wise periodicity, using a fixed length period.

### 2.6.4 Visualization

Visualization is an important mechanism to present the processed time series for further analysis by users. It is also a powerful tool to facilitate the mining tasks like pattern searching, query-by-example, and pattern discovery afterwards. In [36, 37] the authors develop a tool called Time Searcher which is a time series exploratory and visualization tool, so that a user can

retrieve time series by querying. Another time series visualization tool called Viz Tree is proposed [38].

### 2.6.5 Mining time series

Mining is the final goal to discover hidden information or knowledge from either the original or the transformed time series data. Pattern discovery is the most common mining task and the clustering method is the most commonly method. Other time series data mining tasks include classification, rule mining and summarization.

- ***Pattern discovery and clustering***

It is a non-trivial task to discover interesting patterns within time series data, which include frequently appearing [39] and surprising patterns [40] with applications in many domains [41]. In particular, [42] presents a support vector regression (SVR)-based online novelty detection algorithm.

For the problem of time series pattern discovery, a common group of techniques being employed is distance-based clustering [43]. In general clustering procedure, the winner cluster is found and its center is updated accordingly in each iteration. The initial cluster centers can be chosen in various ways. The number of cluster is a critical parameter to be determined. It can be fixed beforehand or can vary during the clustering process. The clustering procedure is finally terminated when the number of iteration exceeds the maximum allowed number of iterations or convergence.

While patterns can be directly discovered from time series, a major problem is that time series data mostly increase linearly with time. This causes storage needs to increase rapidly and slows down the pattern discovery process exponentially. A neural clustering method, the self-organizing map (SOM) [44], is used for pattern discovery with variations adopting emergent feature maps [45, 46].

In [47] the authors adopt the fuzzy c-means (FCM) algorithm for short and unevenly spaced time series clustering. They propose a similarity of short time series based on shapes, which is formed by the relative change of amplitude and the temporal information. A clustering-based method to discover climate indices that represent regions with relatively homogeneous behavior is presented in [48]. Another approach [49] focus on the problem of clustering time series of different lengths, using mixtures of autoregressive moving average (ARMA) models and expectation-maximization (EM) algorithm. In [50], clustering data are derived from ARMA models, using k-means and k-medoids algorithms.

Hidden Markov model (HMM) is a common model-based algorithm adopted in time series clustering [51]. HMMs are defined as stochastic generalizations of finite-state automata, where both transitions between states and generation of output symbols are governed by probability distributions. On the other hand, the discovery of periodic patterns forms another common focus for pattern discovery. In [52], the strategy searches for weak periodic signals using

autocorrelation function and fast Fourier transform (FFT) with no period length are known in advance.

Cluster analysis is also applied on sliding window from the time series for grouping related subsequence patterns that are dispersed along the time series. Such clustering methods seek for a special type of local structure, namely for grouping tendencies in the data. An unfolding (subsampling) preprocessing method is used in [53] before the subsequent SOM clustering. A structure-aware algorithm is proposed by [54] to find exact motifs in massive time series databases.

### ➢ *Classification*

Classification is a traditional data mining task. In the time series domain, special treatment must be placed due to the nature of the data. In [55], authors propose to classify time series data based on combining local properties or patterns in the time series. Moreover, a representation method using wavelet decomposition is suggested in [56]. This method can automatically select the parameters for the classification task. On the other hand, researchers have also focused on customizing or developing classifiers for time series data. For instance, [57] presents a signal classification approach based on modeling the dynamics of a system as they are captures in a reconstructed phase using Gaussian Mixture models of time domain signatures.

### ➢ *Rule discovery*

Rule mining is another typical task in the field of data mining. Association rule mining [58] is one of the most well-known strategies. However, it is mainly focused on symbolic items and many researchers propose new or modified algorithms for rule mining in the context of time series data. Decision tree is another common approach for rule mining. In [59], the preprocessing step to discover interesting rules from the medical time series data is firstly introduced.

### ➢ *Summarization*

Some researchers focus on summarizing and describing the times series data for analysis, mining or prediction. An automated identification of significant qualitative features (interesting patterns) in complex objects is proposed in [60]. Clustering techniques are adopted to summarize and produce a compact description of salient and their relations. Furthermore, [61] proposes to use the fuzzy quantifier to present a linguistic summarization on the trends of time series which the trends are identifies by PLR.

# References


1 Tak-chung Fu, “A review on time series mining”, Engineering Applications of Artificial Intelligence, 24 (1), 2011, pp. 164-181.

2 K.J. ASTROM, “On the choice of sampling rates in parametric identification of time series Information Sciences”, 1 (3) (1969), pp. 273–278.

3 Yi, B., Faloutsos, C., 2000. “Fast time sequence indexing for arbitrary Lp norms”. In: Proceedings of the 26th International Conference on Very Large Data Bases, pp. 385–394.

4 E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, “Dimensionality reduction for fast similarity search in large time series databases Journal of Knowledge and Information Systems”, 3 (3) (2000), pp. 263–286.

5 S. Lee, D. Kwon, S. Lee, “Dimensionality reduction for indexing time series based on the minimum distance Journal of Information Science and Engineering”, 19 (2003), pp. 697–711.

6 Keogh, E., Chakrabarti, K., Mehrotra, S., Pazzani, M., 2001a. “Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data”, pp. 151–163.

7 Shatkay, H., Zdonik, S., 1996. “Approximate queries and representations for large data sequences”. In: Proceedings of the 12th IEEE International Conference on Data Engineering, pp. 536–545.

8 Chung, F.L., Fu, T.C., Luk, R., Ng, V., 2001. “Flexible time series pattern matching based on perceptually important points”. In: International Joint Conference on Artificial Intelligence Workshop on Learning from Temporal and Spatial Data, pp. 1–7.

9 T.C. Fu, F.L. Chung, R. Luk, C.M. Ng, “Representing financial time series based on data point importance”, Engineering Applications of Artificial Intelligence, 21 (2) (2008), pp. 277–30.

10 Man, P., Wong, M.H., 2001. “Efficient and robust feature extraction and pattern matching of time series by a lattice structure”. In: Proceedings of the 10th ACM International Conference on Information and Knowledge Management, pp. 271–278.

11 D.A. Bao, “Generalized model for financial time series representation and prediction” Applied Intelligence, 29 (1) (2008), pp. 1–11.

12 D. Bao, Z. Yang, “Intelligent stock trading system by turning point confirming and probabilistic reasoning” International Journal of Expert Systems with Applications, 34 (1) (2008), pp. 620–627.

13 Jia, P., He, H., Sun, T., 2008. “Error restricted piecewise linear representation of time series based on special points”. In: Proceedings of the Seventh World Congress on Intelligent Control and Automation, pp. 2059–2064.

14 Leng, M., Lai, X., Tan, G., Xu, X., 2009. “Time series representation for anomaly detection”. In: Proceedings of the Second IEEE International Conference on Computer Science and Information Technology, pp. 628–632.

15 Mueen, A., Keogh, E., Bigdely-ShamloN., 2009. “Finding time series motifs in disk-resident data”. In: Proceedings of the 2009 IEEE International Conference on Data Mining, pp. 367–376.

16 Megalooikonomou, V., Li, G., Wang, Q., 2004. “A Dimensionality reduction technique for efficient similarity analysis of time series databases”. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, pp. 160–161.

17 Megalooikonomou, V., Wang, Q., Li, G., Faloutsos, C., 2005. “A Multiresolution symbolic representation of time series”. In: Proceedings of the 21st IEEE International Conference on Data Engineering, pp. 668–679.

18 P.K. Dasha, M. Nayaka, M.R. Senapatia, I.W.C. Lee “Mining for similarities in time series data using wavelet-based feature vectors and neural networks”, Engineering Applications of Artificial Intelligence, 20 (2) (2007), pp. 185–201.

19 Li, C., Yu, P.S., Castelli, V., 1998. MALM: "A framework for mining sequence database at multiple abstraction levels". In: Proceedings of the Seventh ACM International Conference on Information and Knowledge Management, pp. 267–272.

20 Agrawal, R., Faloutsos, C., Swami, A., 1993a. “Efficient similarity search in sequence databases”. In: Proceedings of the Fourth International Conference on Foundations of Data Organization and Algorithms', pp. 69–84.

21 Janacek, G.J., Bagnall, A.J., Powell, M.A., 2005. "Likelihood ratio distance measure for the similarity between the Fourier transform of time series". In: Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 737–743.

22 Yang, K., Shahabi, C., 2005b. "On the stationarity of multivariate time series for correlation-based data analysis". In: Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 805–808.
23 Yang, K., Shahabi, C., 2005a. "A Multilevel distance-based index structure for multivariate time series". In: Proceedings of the 12th IEEE International Symposium on Temporal Representation and Reasoning, pp. 65–73.
24 J.P. Morrill, "Distributed recognition of patterns in time series data", Communications of the ACM, 41 (5) (1998), pp. 45–51.
25 Berndt, D.J., Clifford, J., 1994." Using dynamic time warping to find patterns in time series". In: AAAI Working Notes of the Knowledge Discovery in Databases Workshop, pp. 359–370.
26 Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., Keogh, E., 2003." Indexing multi-dimensional time-series with support for multiple distance measures". In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 216–225.
27 Keogh, E., Lin, J., Truppel, W., 2003. “Clustering of time series subsequences is meaningless: implications for previous and future research”. In: Proceedings of the Third IEEE International Conference on Data Mining, pp. 115–122.
28 Moon, Y., Whang, K., Loh, W., 2001. “Duality-based subsequence matching in time-series databases”. In: Proceedings of the 17th IEEE International Conference on Data Engineering, pp. 263–272.
29 Faloutsos, C., Jagadish, H., Mendelzon, A., Milo, T., 1997. “A Signature technique for similarity-based queries”. In: Proceedings of the International Conference on Compression and Complexity of Sequences, pp. 2–20.
30 Moon, Y.S., Whang, K.Y., Han, W.S., 2002. “General match: a subsequence matching method in time-series databases based on generalized windows”. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, pp. 382–393.
31 X. Liu, Z. Lin, H. Wang, “Novel online methods for time series segmentation”, IEEE Transactions on Knowledge and Data Engineering, 20 (12) (2008), pp. 1616–1626.
32 Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E., 2008. "Querying and mining of time series data: experimental comparison of representations and distance measures". In: Proceedings of the VLDB Endowment, vol. 1(2), pp. 1542–1552.
33 Das, G., Lin, K.I., Mannila, H., Renganathan, G., Smyth, P., 1998." Rule discovery from time series. In": Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 16–22.
34 J. Abonyi, B. Feil, S. Nemeth, P. Arva, "Modified Gath–Geva clustering for fuzzing segmentation of multivariate time-series Fuzzy Sets and Systems", Data Mining Special Issue, 149 (2005), pp. 39–56.
35 Han, W.S., Lee, J., Moon, Y.S., Jiang, H., 2007. "Ranked subsequence matching in time-series databases". In: Proceedings of the 33rd International Conference on Very Large Databases, pp. 423–434.
36 Keogh, E., Chu, S., Hart, D., Pazzani, M., 2001c." An online algorithm for segmenting time series". In: Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 289–296.
37 Srivastava, A.N., Weigend, A.S., 1996." Improved time series segmentation using gated experts with simulated annealing". In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1883–1888.
38 Lin, E. Keogh, S. Lonardi, "Visualizing and discovering non-trivial patterns in large time series databases", Information Visualization, 4 (2) (2005), pp. 61–82.
39 T.C. Fu, F.L. Chung, K.Y. Kwok, C.M. Ng," Stock time series visualization based on data point importance", Engineering Applications of Artificial Intelligence, 21 (8) (2008), pp. 1217–1232.
40 Keogh, E., Lonardi, S., Chiu, Y.C., 2002b. "Finding surprising patterns in a time series database in linear time and space". In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 550–556.
41 Chan, P., Mahoney, M., 2005. "Modeling multiple time series for anomaly detection". In: Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 90–97.
42 Ma, J., Perkins, S., 2003. "Online novelty detection on temporal sequences". In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 613–618.
43 P.K. Dasha, M. Nayaka, M.R. Senapatia, I.W.C. "Lee Mining for similarities in time series data using wavelet-based feature vectors and neural networks", Engineering Applications of Artificial Intelligence, 20 (2) (2007), pp. 185–201.
44 T. Kohonen, “Self-Oranizing Maps”, Springer, Berlin (1995).
45 A. Ultsch, “Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series”, Kohonen Maps (1999), pp. 33–46.

46 F. Morchen, A. Ultsch, O. Hoos ,"Extracting interpretable muscle activation patterns with time series knowledge mining", International Journal of Knowledge-Based+Intelligent Engineering Systems (2005).
47 Moller-Levet, C.S., Klawonn, F., Cho, K.H., Wolkenhauer, O., 2003. "Fuzzy clustering of short time-series and unevenly distributed sampling points". In: Proceedings of the Fifth International Symposium on Intelligent Data Analysis, pp. 330–340.
48 Steinbach, M., Tan, P.N., Kumar, V., Klooster, S., Potter, C., 2003. "Discovery of climate indices using clustering". In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 446–455.
49 Y. Xiong, D.Y. Yeung, "Time series clustering with ARMA mixtures", Pattern Recognition, 37 (8) (2004), pp. 1675–1689.
50 A. Bagnall, G. Janacek, "Clustering time series with clipped data", Machine Learning, 58 (2–3) (2005), pp. 151–178.
51 Panuccio, A., Bicego, M., Murino, V., 2002. "A Hidden Markov Model-based approach to sequential data clustering". In: the Joint International Association for Pattern Recognition Workshops on Structural, Syntactic and Statistical Pattern Recognition, pp. 734–742.
52 Berberidis, C., Vlahavas, I.P., Aref, W.G., Atallah, M.J., Elmagarmid, A.K., 2002a." On the discovery of weak periodicities in large time series". In: Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 51–61.
53 G. Simon, J.A. Lee, M. Verleysen, "Unfolding preprocessing for meaningful time series clustering", Neural Networks, 19 (6) (2006), pp. 877–888.
54 Mueen, A., Keogh, E., Bigdely-ShamloN., 2009. "Finding time series motifs in disk-resident data". In: Proceedings of the 2009 IEEE International Conference on Data Mining, pp. 367–376.
55 Geurts, P., 2001. "Pattern extraction for time series classification". In: Proceedings of the Fifth European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 115–127.
56 Zhang, H., Ho, T.B., Lin, M.S., 2004. "A Non-parametric wavelet feature extractor for time-series classification". In: Proceedings of the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 595–603.
57 R.J. Povinelli, M.T. Johnson, A.C. Lindgren, J. Ye, "Time series classification using Gaussian mixture models of reconstructed phase spaces", IEEE Transactions on Knowledge and Data Engineering, 16 (6) (2004), pp. 779–783.
58 Agrawal, R., Srikant, R., 2000. "Privacy-preserving Data Mining". In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 439–450.
59 Ohsaki, M., Sato, Y., Yokoi, H., Yamaguchi, T., 2003. “A Rule discovery support system for sequential medical data in the case study of a chronic hepatitis data set”. In: the 14th European Conference on Machine Learning/the Seventh European Conference on Principles and Practice in Knowledge Discovery in Databases Discovery Challenge Workshop, pp. 154–165.
60 Zwir, I., Enrique, E.H., 1999. "Qualitative object description: initial reports of the exploration of the frontier". In: Proceedings of Joint EUROFUSE-SIC99 International Conference, pp. 485–490.
61 J. Kacprzyk, A. Wilbik, S. Zadrozny, "Linguistic summarization of time series using a fuzzy quantifier driven aggregation", Fuzzy Sets and Systems, 159 (12) (2008), pp. 1485–1499.

# 3. BOOTSTRAPPING: A STATISTICAL METHOD TO SEARCH FOR HIDDEN SUBCLASSES

## 3.1 Bootstrap introduction and fundamentals

B. Efron and R.J. Tibshirani said that "Statistics is the science of learning from experience, especially experience that arrives a little bit at a time" [1]. The most challenges question on statistical theory could summarize on three questions:

*1. How should I collect the data?*
*2. How should I analyze the data that I have collected?*
*3. How accurate are my data summaries?*

The last question composes section of the process known as statistical inference. The bootstrap is a recently developed technique for making certain kind of statistical inferences. The cause of bootstrap recently development is that requires modern computer power and techniques to simplify the often intricate.

The basic concept behind the bootstrap is very simple and begins two centuries ago. The bootstrap is a data-based simulation method for statistical inference, which can be used to produce inferences. The origin of its name derives from the phrase to ''pull oneself up by one's bootstrap, widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by Rudolph Erich Raspe. The term ''bootstrap'' used in computer science meaning to ''boot'' a computer from a set of core instructions, though the derivation is similar.

The three basic statistical concept, data collection, summary and inference are published in the New York Times of January 27, 1987. This study attempted to see if small aspirin doses would prevent heart attacks in healthy middle-aged men. A controlled, randomized, double-blind study chosen for the collection of aspirin data. Two groups created for this study, the first of the subjects received aspirin and the second received a control substance or placebo with no active ingredients. The statisticians keeping a secret code of who received which substance. The elaborate precautions of a controlled, randomized, blinded experiment guard against benefits that don't exist, while maximizing the chance of detecting a genuine positive effect.

The summary statistics in the study are below:

Heart attacks (fatal plus non-fatal) subjects

| | | |
|---|---|---|
| Aspirin group | 104 | 11,037 |
| Placebo group | 189 | 11,034 |

The aspirin group has the lower rate of heart attacks. The ratio of the two rates is:

$$\hat{\theta} = \frac{104/11037}{189/11034} = 0.55$$

The subjects which received aspirin have 55% heart attacks as many as placebo subjects.

The point here is not the value of the $\hat{\theta}$. The actually question, what we would like to know, is if the experiment was performed again, will it come out equally trustworthy? Statistical theory comes in to make the following inference: the true value of θ lies in the interval $0.43 \prec \theta \prec 0.70$ with 95% confidence.

Note that

$$,\theta = \hat{\theta} + (\theta - \theta) = 0.55 + [\theta - \hat{\theta}(\omega_0)] \qquad 3.1$$

Where θ and $\hat{\theta}(\omega_0)$ (=0.55) are two numbers. In statistics, we use $\theta - \hat{\theta}(\omega)$ to describe $\theta - \hat{\theta}(\omega_0)$. Since ω cannot be observed exactly, we instead study the fluctuation of $\theta - \hat{\theta}(\omega)$ among all ω. If $\theta - \hat{\theta}(\omega)$ is around zero, we can presume statistically that θ is close to 0.55(= $\hat{\theta}(\omega_0)$). If $P(\omega : \left|\theta - \hat{\theta}(\omega)\right| \prec 0.1 = 0.95$, we claim that with 95% confidence that θ-0.55 is no more than 0.1.

In the aspirin study, there are also track strokes. The results are presented below:

| | strokes | subjects |
|---|---|---|
| Aspirin group | 119 | 11,037 |
| Placebo group | 98 | 11,034 |

For strokes, the ratio of the two rates is

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21$$

This result looks like taking aspirin is actually harmful. The interval for the true stroke ratio θ turns out to be 0.93 < θ < 1.59 with 95% confidence. This interval includes the neutral value θ=1, at which aspirin would be no better or worse than placebo. Statistically, we can result that aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes.

According to the above discussion, we use the sampling distribution of $\hat{\theta}(\omega)$ to develop intervals in which the true value of θ lies on with a high confidence level. The task of data analysis is to find the sampling distribution of the chosen estimator $\hat{\theta}$. Actually in practice it means that quite often we are on finding right statistical table to look up.

These tables are constructed based on the model-based sampling theory approach to statistical inference. In this approach, it starts with the assumption that the data are chosen as a sample from some conceptual probability distribution, f. When f is completely specified, we derive the distribution of $\hat{\theta}$, where $\hat{\theta}$ is a function of the observed data. In attempts to derive its

distribution, those data will be considered as random variables. Uncertainties of our inferences can then be measured. The traditional parametric inference utilizes a priori assumptions about the shape of f. From the above example, we rely on the binomial distribution, large sample approximation of the binomial distribution, and the estimate of θ.

Quite often we need to figure out f alter alternatively. Consider a sample of weights of 27 rats (n=27). The data are:

57,60, 52,49,56,46,51,63,49,57,59,54,59,57,52,52,61,59,53,59,51,51,58,46,53.

The sample mean of these data=54.6667, standard deviation=4.5064 with cu=0.0824. In this part the following question arises, what if we wanted an estimation of the standard error of cu. This would be a nonstandard problem. First we may construct a nonparametric f estimator from the sample data. Secondly we can utilize either Monte Carlo method or large sample method to give an approximation on it.

Alternatively, we would follow a different approach from the above. The following nonparametric bootstrap method which relies on the empirical distribution function. This method applies to the stroke example.

a. Create two populations: The first consisting of 119 ones and 11037-119=10918 zeros and the second consisting of 98 ones and 11034-98=10936 zeros.
b. Monte Carlo Resampling: Choose with replacement a sample of 11037 items from the first population and a sample of 11034 subjects from the second population. Each of these is called a bootstrap sample.
c. Derive the bootstrap replicate of $\hat{\theta}$

$\hat{\theta}^* =$ Ratio of ones in bootstrap sample 1 / Ratio of ones in bootstrap sample 2

Repeat this process (a-c) a large number of times, 1000 times and obtain 1000 bootstrap replicates $\hat{\theta}$.

In this sample example, the standard deviation turned out to be 0.17 in a batch of 1000 replicates that we generated. Also a rough 95% confidence interval is (0.93, 1.60) which is derived by taking the $25^{th}$ and $975^{th}$ largest of the 1000 replicates.

After basic introduction to bootstrap, we summarize to three specific points. First of all, the basic bootstrap approach uses Monte Carlo sampling to generate an empirical estimate of the sampling distribution by drawing a large number of samples of size n from an initial population. Then it calculates the associated value of the statistic $\hat{\theta}$ for each one. An estimator of sampling distribution for the statistic is calculated by relative frequency distribution of these $\hat{\theta}$ values. The accuracy of estimation of sampling distribution, it depends from the size of the sample and increases as the sample is getting larger.

Second, with the bootstrap method, the basic sample is considered as the population and a similar to Monte Carlo procedure, lead it. The process starts with a random selection of a large number of resamples of size n from this original sample with replacement. This means that, even though each resample will have the same number of elements as the original sample, it

could include some of the original data points more than once time. The elements in these resamples vary slightly, and because of that the statistic $\hat{\theta}$, which calculated from each one of these resample, will take on slightly different values.

Third, the central assertion of the bootstrap method is that the relative frequency distribution of these $\hat{\theta}_{F_n}$'s is an estimate of the sampling distribution of $\hat{\theta}$.

## 3.2 Bootstrap method

The bootstrap method introduced in Efron (1979) [2] is a very general resampling procedure for estimating the distributions of statistics based on independent observations. The bootstrap method is shown to be successful in many situations, which is being accepted as an alternative to the asymptotic methods. There are [6] several forms of the bootstrap and several other resampling methods that are related to it, such as jackknifing, cross-validation, randomization tests and permutation tests. Here we will stand to the nonparametric bootstrap. We present the basic bootstrap procedure, according to Efron and Tibdhirani, into the following steps as follows. Assume a random sample size n is drawn an unspecified distribution F:

1. Construct an empirical probability distribution $F_n$ from the sample by placing a probability of 1/n at each point, $x_{1,}x_2,...,x_n$ of the sample. This is the empirical distribution function of the sample, which is the nonparametric maximum likelihood estimate of the population distribution, F.
2. From the empirical distribution function, $F_n$, draw a random sample of size n with replacement. To resample is to sample from the empirical distribution. To get us a little closer to implementing (8) this on a computer we describe this as follows. Label the 5 data points $x_{1,}x_2,...x_5$. To resample is to draw a number j from the uniform distribution on {1, 2,…, 5} and take $x_j$ as our resample value.
3. Calculate the statistic of interest, $T_n$, for this resample, calling $T_n^*$.
4. Repeat the steps 2 and 3 for B times, where B is a large number, in order to create B resamples. The size of B is at least to 1000 when an estimate of confidence interval around $T_n$ is required.
5. Construct the relative frequency histogram from The B number of $T_n^*$'s by placing a probability of 1/B at each point, $T_n^{*1},T_n^{*2},...,T_n^{*B}$. The distribution can now be used to make inferences about the parameter θ, which is to be estimated by $T_n$.

The idea behind bootstrap is to use the data of a sample study at hand as a ‘‘surrogate population’’, for the purpose of approximating the sampling distribution of a statistic. The sample summary is then computed on each of the bootstrap samples (usually a few thousand). A histogram of the set of these computed values is referred to as the bootstrap distribution of the statistic [3]. In bootstrap’s most elementary application, one produces a large number of ‘‘copies’’ of a sample statistic, computed from these phantom bootstrap samples. In continuous, a small percentage, like 100(α/2) % (usually a=0.05), is trimmed off from the lower as well as from the upper end of these numbers. The range of remaining 100 (1-α) % values is declared as the confidence limits of the corresponding unknown population summary number of interest, with level of confidence 100 (1-α) %. The above method is referred to as bootstrap percentile method. For more sample statistics, the sampling distribution of $\hat{\theta}$ for a large sample n, is

shaped with center θ and standard deviation $\left(\alpha/\sqrt{n}\right)$, where the positive number depends from the type of statistic $\hat{\theta}$. This is the known phenomenon as, Central Limit Theorem (CLT) [5].

## ✓ Generation of bootstrap population

➢ *Parametric bootstrap*

Bootstrapping is general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand. Consider that we select a sample S=$\{X_1, X_2, ..., X_n\}$ from a population P=$\{x_1, x_2, ..x_N\}$. Furthermore, imagine that N is very larger than n and that S is either a simple random sample or an independent random sample from P. It will also help initially to think of the elements of the population as a scalar value, but they could just as easily be vectors (i.e., multivariate). Suppose that we are interested in some statistic T=t(S) as an estimate of the corresponding population parameter θ=t (P). For θ could be a vector of parameters and T the corresponding vector of estimates, but for simplicity considers that θ is a scalar [6]. An ordinary approach to statistical inference is to make assumptions about the structure of the population (e.g., an assumption of normality) and along with the stipulation of random sampling, to use these assumptions to derive the sampling distribution of T, on which classical inference is based. In certain instances, the exact distribution of T may be intractable and so we instead derive its asymptotic distribution. This known approach has two potentially important deficiencies:

1. If the assumptions about the population are wrong, then the corresponding sampling distribution of the statistic may be seriously inaccurate. On the other hand, if asymptotic results are relied upon, these may not hold to the required level of accuracy in a relatively small sample.
2. The approach requires sufficient mathematical prowess to derive the sampling distribution of the statistic of interest. In some cases, such a derivation may be prohibitively difficult.

➢ *Nonparametric bootstrap*

The nonparametric bootstrap allows us to determine the sampling distribution of a statistic empirically without making assumptions about the form of the population, and without deriving the sampling distribution explicitly. The essence of the nonparametric bootstrap idea is as follows: We select randomly a sample of size n form among the elements of S, sampling with replacement. Suppose we call the resulting bootstrap sample $S_1^* = \{X_{11}^*, X_{12}^*, ..., X_{1n}^*\}$. We choose to sample with replacement, because we would otherwise simply reproduce the original sample S. In effect, we are treating the sample S as an estimate of the population P; that is each

element $X_i$ of S is selected for the bootstrap sample with probability 1/n, imitating the original selection of the sample S from the population P. We repeat this procedure a large number of times, R, selecting many bootstrap samples; the bth such bootstrap sample is denoted $S_b^* = \left\{ X_{b1}^*, X_{b2}^*, ..., X_{bn}^* \right\}$.

## ✓ Estimation of the bootstrap statistic (the optimal value of the statistic based on the bootstrap population)

One way is to estimate it as the mean of all estimates in the bootstrap process. We propose a different approach that focuses on the distribution of the statistic itself. In other words, by considering the samples of the statistic across the bootstrap process as meta-data, we can design a second computational process with the aim to estimate the mode peak(s) of this newly created distribution. One exemplar approach towards this direction is implemented by the Mean Shift scheme.

At this point we can consider several accuracy issues in relation to the estimation of the bootstrap value of the statistic of interest. Initially, we compute the statistic T for each of the bootstrap sample; that is $T_b^* = t(S_b^*)$. Then the distribution of $T_b^*$ around the original estimate T is analogous to the sampling distribution of the estimator T around the population parameter θ. For example, the average of the bootstrapped statistics,

$$\overline{T}^* = \hat{E}^*\left(T^*\right) = \frac{\sum_{b=1}^{R} T_b^*}{R} \qquad 3.2$$

Estimates the expectation of the bootstrapped statistics; then $\hat{B}^* = \overline{T}^* - T$ is an estimate of T, that is, T-θ. Similarly, the estimated bootstrap variance of $T^*$,

$$\hat{V}^*\left(T^*\right) = \frac{\sum_{b=1}^{R}\left(T_{b=1}^{R} - \overline{T}^*\right)^2}{R-1} \qquad 3.3$$

estimates the sampling variance of T.

The random selection of bootstrap samples is not an essential aspect of the nonparametric bootstrap: At least in principle, we could enumerate all bootstrap sample of size n. Then we could calculate $E^*\left(T^*\right)$ and $V^*\left(T^*\right)$ exactly, rather than having to estimate them. The number of bootstrap sample, however, is very large unless n is $tiny^2$. There are, therefore, two sources of error in bootstrap inference:

1. The error induced by using a particular sample S to represent the population.

2. The sampling error produced by failing to enumerate all bootstrap samples. This source of error can be controlled by making the number of bootstrap replications R sufficiently large.

## 3.3 Bias correction by bootstrap

Suppose $\hat{\theta}$ is a sample estimator of θ based on a random sample of size n, i.e. $\hat{\theta}$ is a function of the data $(X_1, X_2, ..., X_n)$. In order to estimate the standard error of $\hat{\theta}$ we must follow the next path:

In the beginning, we should draw many bootstrap samples N, of same size n to the initial sample $\left(\theta_1^*, \theta_2^*, ..., \theta_n^*\right)$. In continuous, we calculate the standard deviation or standard error that corresponds to a standard deviation and that is actually an estimate of this standard deviation. One defines $SE_B(\hat{\theta}) = \left[ (1/N) \sum_{i=1}^{N} \left(\theta_i^* - \hat{\theta}\right)^2 \right]^{1/2}$ following the idea of bootstrap: replace the population by the empirical population.

The mean of sampling distribution of $\hat{\theta}$ often is different from θ, usually by an amount= $c/n$, for large n. In statistics, this one is described as

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \approx O(1/n) \qquad 3.4$$

The bootstrap approximation to this bias is as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\theta_i^* - \hat{\theta} = \hat{Bias}_B\left(\hat{\theta}\right) \qquad 3.5$$

Where $\theta_i^*$ are bootstrap copies of $\hat{\theta}$, as defined in the earlier subsection? The bootstrap bias corrected estimator is $\hat{\theta}_c = \hat{\theta} - \hat{Bias}_B\left(\hat{\theta}\right)$.

## 3.4 Bootstrap Confidence Intervals

There are several approaches to constructing bootstrap confidence intervals [7]. The normal-theory interval assumes that the statistic T is normally distributed (which is approximately the case for statistics in sufficiently large samples), and uses the bootstrap estimate of sampling variance and perhaps of bias to construct a 100(1-a)-percent confidence interval of the form [7].

$$\theta = \left(T - \hat{B}^*\right) \pm z_{1-a/2} \hat{SE}^*(T^*) \qquad 3.6$$

Here, $\hat{SE}(T^*) = \sqrt{\hat{V}^*(T^*)}$ is the bootstrap estimate of the standard error of $T$ and $z_{1-a/2}$ is the $1-a/2$ quantile of the standard-normal distribution.

An alternative approach, is the bootstrap percentile interval [7] This method is very popular because its simplicity and natural appeal. Suppose one settles for 1000 bootstrap replications of $\hat{T}$, denoted by $\left(T_1^*, T_2^*, ..., T_{1000}^*\right)$. After ranking from bottom to top, let us denote these bootstrap values as $\left(T_{(1)}^*, T_{(2)}^*, ..., T_{(1000)}^*\right)$. We summarize the empirical quintiles of $T_b^*$ to form a confidence interval for θ as follows:

$$T_{(lower)}^* < \theta < T_{(upper)}^* \qquad 3.7$$

where $T_{(1)}^*, T_{(2)}^*, ..., T_{(R)}^*$ are the ordered bootstrap replicates of the statistic; lower $= \left[\left(R+1\right)a/2\right]$; upper $= \left[\left(R+1\right)\left(1-a/2\right)\right]$; and the square brackets indicate rounding to the nearest integer. For example, if a=0.05, corresponding to a 95-percent confidence interval, and R=999, then the lower=25 and upper=975.

## 3.5 Determination of number of repetitions

The choice of correct number of bootstrap repetitions is the most of the times in our judgment. Kesar Singh and Minge Xie [4] are making a crude recommend for the size N of different bootstrap samples, $N = n^2$ unless $n^2$ is too large.

Later in 2000, Donald W. K. Andrews and Moshe Buchinsky [9] suggested a three-step Method for choosing the number of bootstrap iterations B for bootstrap standard errors, confidence intervals, confidence regions, hypothesis tests, p-values and bias correction. The choice of $\omega_1$ for above reasons, is based on asymptotic distribution of $T^*$ as $n \to \infty$. The asymptotic distribution does not require being close to the finite sample distribution for the three-step method to work well. The cause is that the initial value of $\omega_1$ is used only to generate an initial value of B that is used, in turn, to obtain an improved value of $\omega$ that reflects the finite sample distribution of $\hat{\theta}^*$ or $T^*$.

Let $\text{int}(a)$ signify the smallest integer greater that or equal to $\alpha$. The three-step method is as follows:

*Step 1*: $$B_1 = \text{int}\left(10,000 z_{1-\tau/2}^2 \omega_1 / pdb^2\right) \qquad 3.8$$

Or, if $\hat{\lambda}_B$ is an $\alpha$ or $1-\alpha$ sample quantile, compute

$$B_1 = \alpha_2 h_1 - 1, \qquad 3.9$$

where $a = a_1/a_2$ and $h_1 = \text{int}(10,000 z_{1-\tau/2}^2 \omega_1 / \left(pdb^2 a_2\right))$

*Step 2*: Simulate $B_1$ bootstrap samples $\left\{ X_b^* : b = 1, ..., B_1 \right\}$ and compute an improved estimate $\hat{\omega}_{B_1}$ of $\omega$ using the appropriate formulae, with $B$ replaced by $B_1$.

*Step 3*: Compute

$$B_2 = \text{int}(10,000 z_{1-\tau/2}^2 \hat{\omega}_{B_1} / pdb^2) \qquad 3.10$$

or, if $\hat{\lambda}_B$ is an $\alpha$ or $1-\alpha$ sample quantile, compute

$$B_2 = a_2 h_2 - 1, \text{ where } h_2 = \text{int}\left(10,000 z_{1-\tau/2}^2 \hat{\omega}_{B_1} / \left( pdb^2 a_2 \right)\right).$$

Take the desired number of bootstrap repetitions to be $B^* = \max\left\{B_2, B_1\right\}$.

Note that Steps 2 and 3 could be with little additional computational burden by replacing $B_1$ in Step 2 by $B_2$, replacing $B_2$ in Step 3 by $B_3$ and taking $B^* = \max\left\{B_3, B_2, B_1\right\}$. In some cases, this may lead to finite sample properties that are closer to the asymptotic properties of the three-step procedure.

In our bootstrap process, we preferred to utilize a number of 500 repetitions as the number of sample, that we random select in every repetition. We develop that in next chapter.

## 3.6 Comparison with other resampling techniques

Statistical resampling methods have become feasible for parametric estimation, hypothesis testing and model validation now that the computer is a ubiquitous tool for statisticians. The bootstrap is similar to earlier techniques which are also called resampling methods and we will analyze and compare many of them with bootstrapping below:

1. Jackknife
2. Cross-validation
3. Permutation test

### 3.6.1 Jackknife

The Jackknife was proposed by M.H. Quenouille in 1949 and later refined and given its current name by John Turkey in 1956 [10]. Quenouille originally developed the method as a procedure for correcting bias (A statistical sampling or testing error caused by systematically favoring some outcomes over others). Later, Turkey develops its use to construct confidence limits for a large of estimators. The main difference to bootstrap is that jackknife samples without replacement, instead of bootstrap method, sampling with replacement.

In many statistical situations is impractical or very difficult to calculate good estimators or find those estimators' standard errors. So, Jackknife is a technique to obtain reliable statistical

estimators. As we mentioned before, Quenouille (1949) introduced a method for reducing the bias correction of a serial correlation estimator based on splitting the sample into two half-sample. Later in a paper of 1967, he generalized this idea and preceded this idea into splitting the sample into $g$ group of size h each, $n = gh$ and explores its general applicability.

Jackknife Samples are selected by taking the original data vector and deleting one observation from the set. In this way, there are $n$ unique Jackknife samples, and the $ith$ Jackknife sample vector is defined as [11]:

$$X_{[i]} = \{X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots, X_{n-1}, X_n\} \qquad 3.11$$

Let $\hat{\theta}_{(i)} = s\left(X_{(i)}\right)$ be the $ith$ Jackknife replication of $\hat{\theta}$. The jackknife estimate of standard error defined by

$$s\hat{e}_{jack} = \left[\frac{n-1}{n}\sum_i \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right]^{1/2} , \qquad 3.12$$

Where $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)} / n$.

The $n-1/n$ factor in the formula above looks similar to the formula for the standard error of the sample mean, except that there is a quantity $(n-1)$ included in the numerator.

The Jackknife is accuracy with linear statistics, but it fails to accurate estimation for non-smooth and nonlinear cases. Because of this, it is less general technique that bootstrap. The different way that explores sample variation, is also a difference from bootstrapping. In addition, Jackknife samples without replacement and yields the same result each time, while bootstrap involves sampling with replacement. Bootstrap overshadows Jackknife because is a more thorough procedure in the sense that it draws many more subsamples that the other and has not bound to theoretical distributions. Through simulations, it is found that the bootstrap method provided less biased and more consistent results than the Jackknife method does. When the purpose of resample g is to determine how each sub-sample affect any model, Jackknife is the right option.

### 3.6.2 Cross Validation

Cross-validation is statistic implementation, to predict the performance of statistical model [11]. The basic method is the partitioning data into a sample of data into sub-samples. In this way, the initial analysis is conducted on a single sub-sample (training data), while further sub-samples (the test or validation data) are retained ''blind'' in order for subsequent use in confirming and validating the initial analysis. For example, the predictive accuracy of a model can be measured by the mean squared error (MSE) on the test set. This will generally be larger than the MSE on training set, because the test data were not used for estimation.

In contrast to bootstrap, cross- validation is not a resampling technique. It is more implementable for large amounts of data, while bootstrap requires small amount of data. Cross-validation is extremely useful in fields such as, data mining and artificial intelligence. Even though the principles of cross-validation, Jackknife and bootstrap are very similar, the bootstrap overshadows the others for it is a more thorough procedure in the sencw that it draws many more sub-samples than the others.

### 3.6.3 Permutation test

The permutation test is also known as the randomization exact test. R. A. Fisher (1935) was the first who developed this method. The basic idea behind permutation technique is to generate a reference distribution by recalculating a statistic for many permutations of data. [12]. Later Fisher wrote that ''the statistician does not carry conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method''.

Randomization exact is a test procedure in which data are randomly re-assigned so than an exact p-value is calculated based on the permutated data. In permutation test such as any other test statistic, there is a null Hypothesis, Ho. The null Hypothesis that is followed here is that some of data are exchangeable. So, we permute the data by shuffling their labels of treatments and then calculate our test statistic on each permutation. The collection of test statistic from the permuted data constructs the distribution under Ho.

# References


1. B. Efron, R. J. Tibshirani, “An Introduction to the Bootstrap”, New York: Chapman and Hall, 1993.

2. B. Efron, “Bootstrap Methods: Another Look at the Jackknife”, Annals of Statistics, 7, pp 1:27, 1979.

3. B. Efron, R.J. Tibshirani, “An introduction to the bootstrao”, Monographs on Statistics and Applied Probability, No 57, Chapman and Hall, London, pp 436.

4. Kesar Singh, Minge Xie, “Bootstrap: A statistical Method”.

5. Kesar Singh, “On Asynphotic accuracy of Efron’s bootstrap”, Ann. Stat. Vol 9, pp 1187:1195, 1981.

6. John Fox, “Bootstrapping Regression Models” January 2002.

7. P. Hall, “Bootstrap Confidence intervals in nonparametric regression”, Ann. Stat, Vol 20, pp 695:711, 1992

8. Jeremy Orloff, Jonathan Blomm, “Bootstrap Confidence Intervals”, Class 24, 18 May, Spring 2014.

9. Donald W. K. Andrews, Moshe Buchinsky, “Evaluation of a three-step method for choosing the number of bootstrap repetitions”, Journal if Economists, Vol 103, pp 345:386, 2001.

10. M. Quenouille, “Approximate tests of correlation in time series”, Journal of the Royal Statistical Society, Soc., Series, B, Vol 11, pp 18:84, 1949.

11. Avery I McIntosh, “The Jackknife Estimation Method”.

12. Michael D. Ernst, “Permutation Methods: A basis for Exact Inference”, Statistical Science, Vol 19, No 4, pp 675:685, 2004.

# 4. MATERIALS AND METHODS

## 4.1 Experimental data

In this study, the experimental data represent intensity curves of backscattering light (IBSL) versus time, within the duration of the AW effect, for a selected spectral band in the area of all image pixels. 3D images of the cervical tissue are formulated, with the third dimension representing time. When light is transmitted over tissue sections, absorption or scattering form their components occurs.Furthermore, the application of acetic acid progressively changes the optical properties of the abnormal epithelium, since the disease has caused structural and functional alterations, so that it equally scatters all the incident wavelengths instead of sustaining its transparency, yielding the differentiation of the intensity and the spectral characteristics of the backscattering light from the cervix. Base on this fact, quantitative assessment of the phenomenon could be achieved via the measurement of IBSL as function of both time and wavelength on any spatial point of the area of interest.

Different states of cervical tissue impose different response to acetic acid application. For example, high-grade cervical intraepithelial neoplasia (CIN) sustains an almost instant response within about 50-60 seconds and the acetowhitening effect is then slowly reversed, finally disappearing after about 40 seconds. In the case of low-grade CIN, the onsetof white is commonly delayed because the acid must penetrate into the lower parts of the tissue. Dynamic Spectral Imaging (DSI) has given encouraging results for the quantitative measurement and mapping of the dynamic light-scattering characteristics of epithelium. Initially, it is necessary to determine, through spectral analysis, the wavelength range in which the maximum difference of light scattering is measured among tissue segments with acetic affected (tissue whitening) or unaffected epithelium (no response or no marker applied). In parallel, the contrast between these areas and the signal-to-noise in the recorder IBSL versus time curves be maximized. Previous studies on Dynamic Spectral Imaging (DSI) [1] confirm the utility of a wavelength in the range of 510-540 nm for this particular application.

The Dynamic Contrast Enhanced Optical Imaging (DCE-OI) device [2], equipped with an optical head adjusted to a mechanical basis and connected to a vaginal speculum for efficient field-of-view during the examination procedure, is utilized for the acquisition of image samples. Tissue imaging is achieved via the appropriate setup of a 35 mm lens, a 1024x768, 8 bit/channel color CCD sensor, a white Light Emission Diode (LED) and light collimating optics to ensure uniform tissue illumination. The optical head is configured to capture images from a 23mmx20mm tissue area, including the entire transformation zone of the cervix periodically recording the aceto-whitening effect. The beginning of the image capturing is triggered by an Acetic Acid applicator in order to ensure synchronization between the acquisition scheduling and the phenomenon. A reference image is captured before the application of the AA solution, which is followed by an automatic image capturing of images during the evolution of the AW phenomenon for the interval of 4 minutes, time window within which the biomedical effect has

evolved and attenuated. From captured image stack, diffuse reflectance curves arecalculated for every image pixel, expressing the temporal characteristics of the AW phenomenon normalized to the corresponding reference level. The illustration of the biological phenomenon on a representative image sample is depicted in Figure 5. The intensity of the backscattered light as a function of time is extracted from the green channel of the captured image series, due to the great dynamic range and high S/N ratio of the specific RGB component. In this essence, each image stack can be viewed as 3-dimensional MxNxK matrix, where M and N are the spatial dimensional of each image and K is the number of captured images (directly connected to the image capturing method). This procedure is depicted in Figure 6.

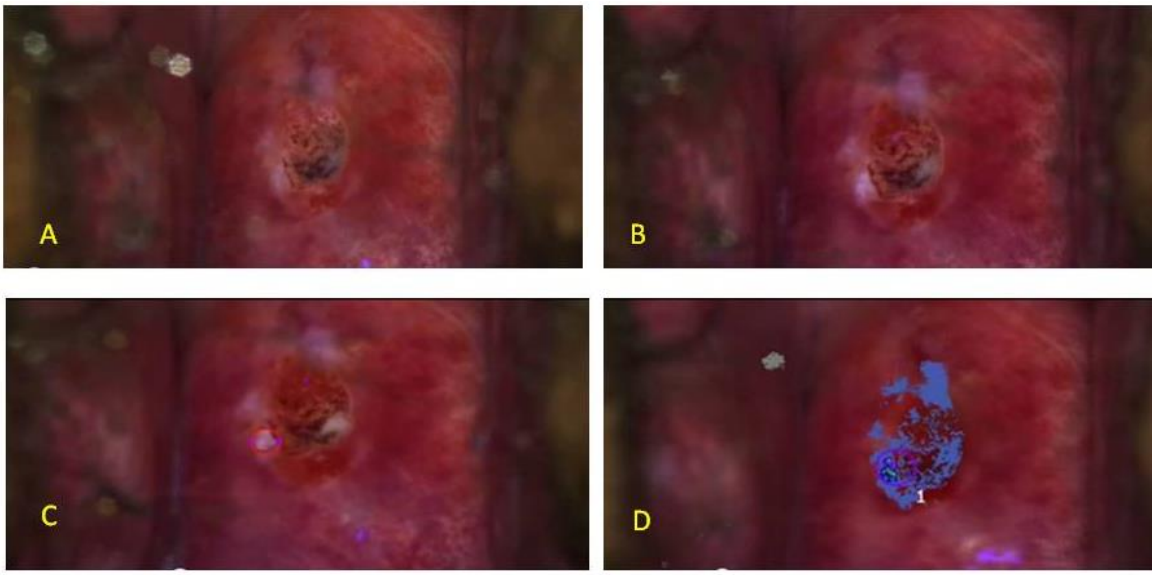


Figure 5. Colposcopic assessment example: (A) Initialization of the aceto-whitening effect taking place on the cervix- (B) Evolution of the aceto-whitening on a latter time slice-(C) Segment of extreme aceto-whitening indicated by the small red circle-(D) Pseudo-colored image map representing the suspicious abnormal tissue areas.

Clinical cases from the University Hospital of Crete were included for the needs of this study. Informed consent was obtained from each clinical case, while the review board of the University Hospital of Crete approved the study. The group underwent the examination procedure and for each individual the IBSL-versus-time plot for each spectral band was automatically recorded and displayed for any selected image area. The samples were selected from a group of woman expressing normal and abnormal Pap smear and ranging in age from 35 to 44 years (mean age 42 years). For each case with abnormal epithelium status was detected, biopsy sampling from epithelium and evaluation was performed by a medical expert, producing a final set of 7 reference curves, representing the seven possible cervical tissue types, covering all the stages of disease as well as the normal case: Normal tissue, Inflammation , HPV, CIN-I, CIN-II, CIN-III and Cancerous tissue. Although the epithelial tissues constantly undergo several

regeneration activities, they sustain the highest risk of developing neoplastic lesions. In fact, 85 out of every 100 cancers are epithelial.

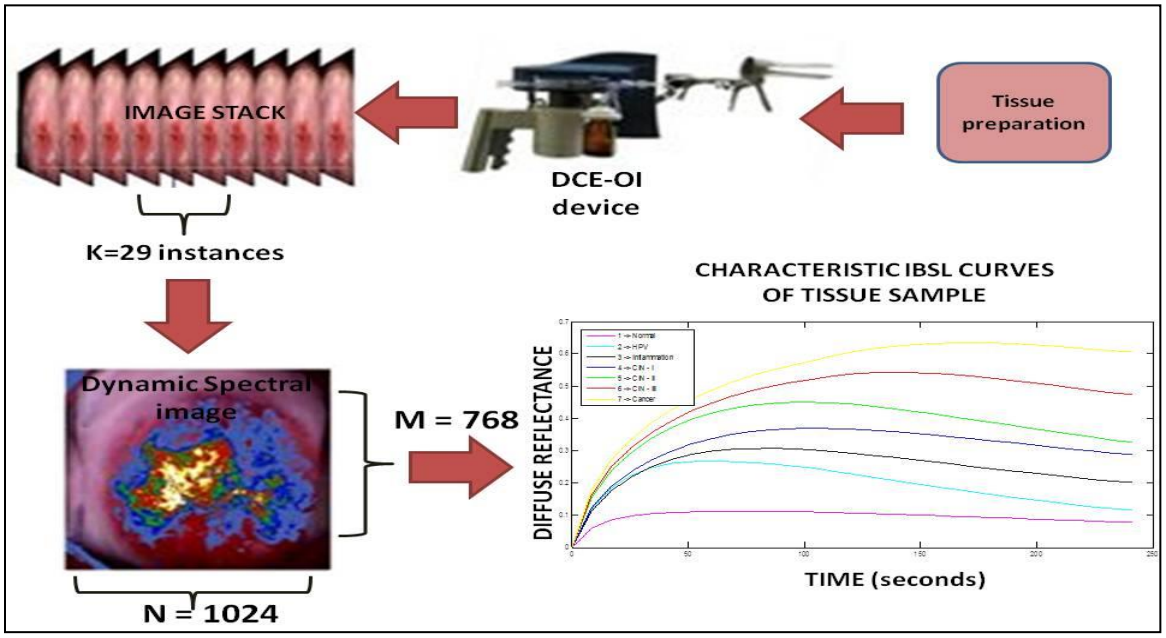


Figure 6. Procedure of converting image series captured during cervical examination to characteristic tissue IBSL vs time curves.

In our experiments, we process 768x1024 biomedical images (M=768, N=1024), recorded at 29 (K=29) distinct time moments time moments to monitor the aceto-whitening effect. Each IBSL curve expresses intensity of backscattered light at 29 time instances. Two data sets are available for analysis aiming at both algorithmic design and evaluation. The first data set contains 497 samples in the form of a (497,29) matrix, for which we also have information regarding the cellular state of the cancer through actual biopsy (71 instances for each of seven cancer states). We also given a set of 7 reference curves in the form of a (7, 29) matrix reflecting representative responses of the seven clinical states of particular interest. These seven curves, sustaining differentiation in both shape and amplitude, have been specified through clinical validation of representative biopsies and reflect the clinically accepted typical responses of the 7 distinct cancer types. Because of that, there exists always a suspicion on the extent to whih this gold standard is 100% valid. Through our algorithmic scheme, we aim to extract and utilize information from the data distribution in order to confirm the knowledge originating from clinical analysis and testing. The second dataset expresses (100.000, 29) samples of IBSL versus time measurements without any further information on the molecular states of cancer. The algorithmic aim of our study is to organize the latter dataset into compound classes that resemble and reproduce the 7 clinical cancer stages. A validation stage after the clustering incorporates the former dataset in order to assess the validity of the results of the proposed algorithmic scheme based on a limited number of known samples. Finally, our aim of clinical interest is to assess any subclasses that can be indentified from the large number of samples in the second dataset, giving rise to potentially interesting substages of the disease.

## 4.2 Proposal methodology

- ***Compatibility of clinical trends with measured responses***

A set of clinical-reference cluster centers Figure 7, which reflects the expert knowledge of the clinical expectation of the typical response curves in each stage of the pathology, is provided in this thesis. Also a dataset of 497 clinical cases is given, experimentally labeled by a medical expert based on biopsy. In this section we aim to examine how well the seven reference curves, even though obtained from clinical experience, can represent the testing set of 497 data samples. The experimental procedure for obtaining the input dataset along with the theoretical back beyond clustering and center refinement was presented in 4.1.

We proceed to group the 497 cases based on the minimum Euclidean distance from 7 clinical reference curves The result is depicted in Figure 8 & Figure 9. The first figure illustrates the agreement of confusion matrix among the assigned and computed labels and the second figure presents the class distributions. It is obvious that the clinical knowledge does not fit well the experimental dataset, since heavy mixing of classes and sample dispersion are observed. Thus, the need arises for appropriate adjustments of data partitioning based on both clinical and experimental evidence.

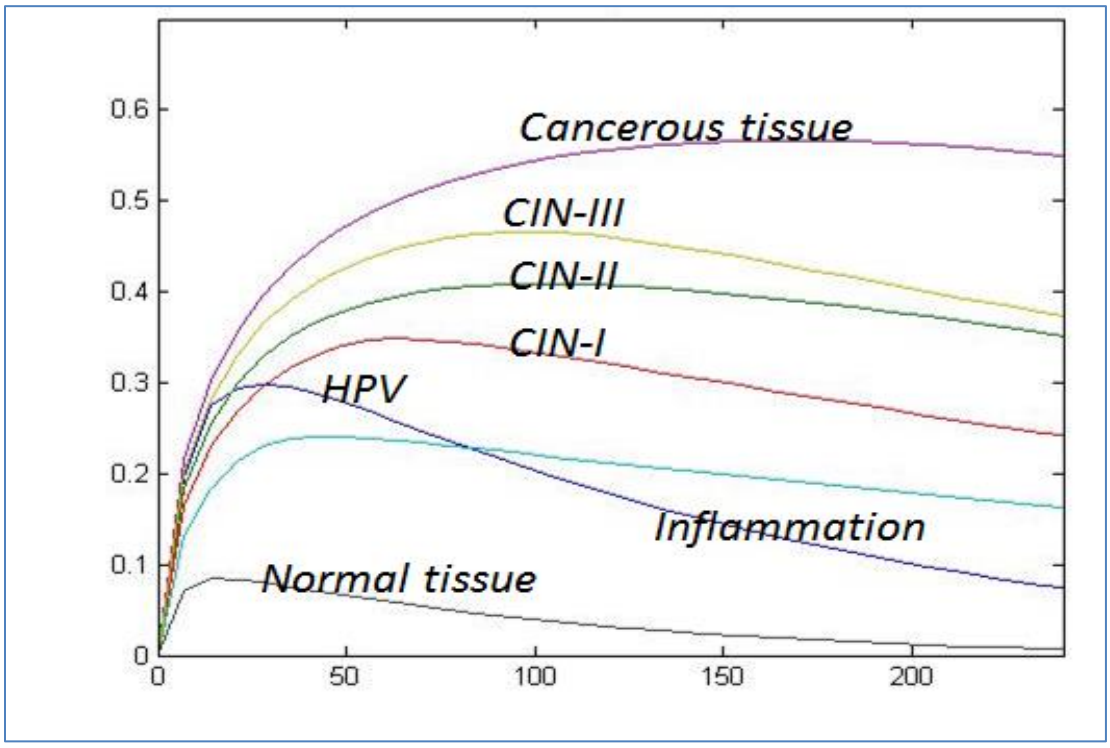


Figure 7. Qualitative representation of the 7 clinical reference centroids according to clinical trends.

*CLASSIFICATION RESULTS ( derived labels per class)*

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 9 | 62 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 0 | 10 | 61 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 63 | 8 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 0 | 40 | 31 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 0 | 13 | 58 | 71 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 71 | 71 |
| count | 71 | 9 | 72 | 124 | 48 | 44 | 129 | |

Figure8. Quantitative results of experimental dataset based on the minimum Euclidean distance represented by confusion matrix

Clustering constitutes an unsupervised machine learning approach aiming to organize the available data into compact classes according to some notion of similarity [3]. It sustains two fundamental properties:

a) The homogenous groups (clusters) are extracted in such a way that items within a cluster are more similar to each other than they are to members of the other groups.

b) Representative patterns within the data are calculated without any prior knowledge on it, since data similarity is considered enough to describe compact classes in a feature space.

Over the years, despite the emergence of numerous clustering approaches, the k-means algorithm is still considered as one of the most competitive and widely used ones for grouping populations [4]. The basic traits of k-means, simplicity, time-efficiency and direct combination with alternative methods in larger systems are considered the principal advantages of k-means clustering. However, there is limitation, that the number k of dominant clusters along with the shape of the data distribution must be known. The main idea behind this technique lies on the

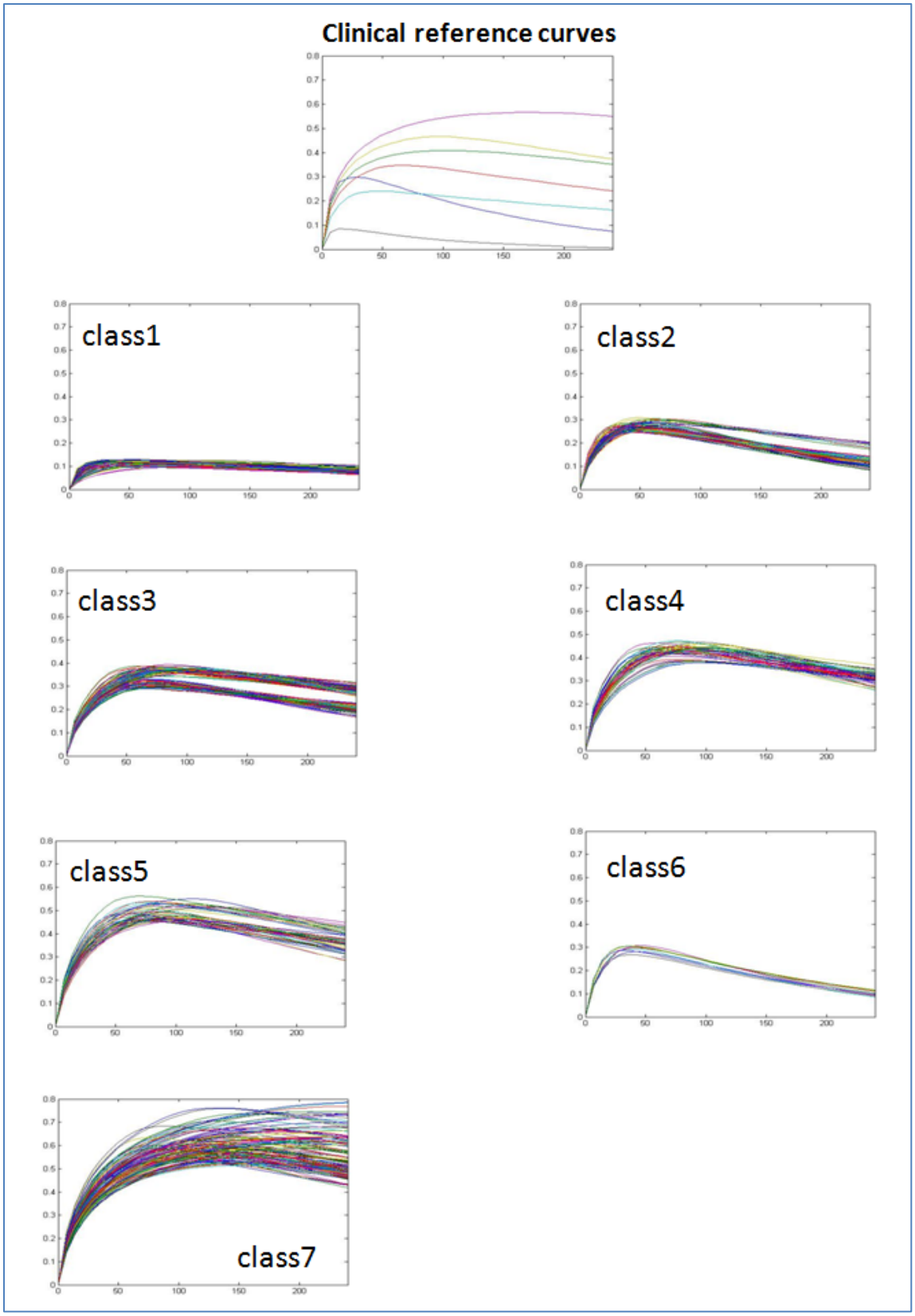


Figure 9. Data grouping of the experimental dataset based on the minimum Euclidean distance from the clinical reference curves.

initial selection of k “starting points”/seeds as cluster centers and their sequential update based on the sample population in the average associated with each cluster as opposed by closer similarity distance.

Numerous alternatives of the classic k-means approach can be found in international bibliography. In [5], the dip-means algorithm was introduced as a promising and robust solution to a key problem in data clustering i.e. that of estimating the number of candidate clusters. This learning was achieved via an incremental method, which initially performs a local search clustering, initialized by a model of k clusters. Then, it performs a novel statistical test on the empirical density distribution of the data for unimodality, in order to decide whether a data subset contains multiple cluster structures. Finally, through a divisive procedure, the selected data subsets are split into two clusters, providing two centers depending on the established decision and control parameters. A novel approach to calculate several clustering solutions for exploratory data analysis, instead of one to provide poor guidance to data analysts was introduced in [6], based on the idea that alternative clustering results may reside in different subspaces. The proposed methodology simultaneously discovers these subspaces along with the corresponding clusters, via an optimization procedure that fuses knowledge for improved cluster quality and novel knowledge related to previously estimate clustering solutions.

Another direction to differentiate from the classic k-means framework yields in the experimentation on the distance metric to be utilized for the assignment of data points to cluster centers. The k-harmonic means algorithm (KHM) is a method similar to k-means (KM) that arises from a different objective function, the harmonic mean of the distance from each data point to all centers [7]. This technique proves promising in finding efficient clustering solutions rapidly, gaining more influence to data points that are not well-modeled by the clustering solution. A good (low) score of the harmonic mean for each data point is achieved when this is close to any of the centroids.

As mentioned above, this thesis attempts to overcome several limitations of self-organized clustering approaches pertaining to stabilization and generalization of the algorithms. Furthermore, it aims at combining expert knowledge with structural information from the data in order to make the data distributions more compatible with the clinically accepted response. We built our approach in four scenarios of incremental complexity on the assumptions and the strategic objectives aimed. The first one considers the stabilization of the k-means algorithm to make it robust over the initialization and the number of classes. This scheme, however does not utilize the clinical knowledge (or the clinically accepted classes with their centers). The second scenario attempts to combine both clinical and data information by building an incremental scheme where the centers initialized by expert knowledge are updated and incrementally corrected via the data distributions. Even though the above schemes achieve better agreement on the structure conveyed by the data and the expert classifications, there is still certain mismatch that questions the validity of the metric used for comparisons of data curves. Thus, we examine in the third scenario several combinations of measures and propose a novel metric stemming from such a joint consideration, which places importance in both the magnitude and shape of data. Finally, the fourth scenario examines the generalization ability of bootstrapping

techniques by considering the clustering of big data in smaller subsamples and the combination of results towards the total dataset.

### 4.2.1 Scenario 1: Data Self-Organization through Resampling and Clustering Approaches without Prior Knowledge

K-means clustering aims at self-organization of a population based on similarity of data vectors within with class or with a single reference vector for each class. Since it also considers pair similarities without information of the entire distribution, it suffers from initialization problems that affect algorithmic stability, as well as from appropriate representation of the classes of interest (number of classes and distribution of each class) that often degrade its generalization capabilities. The number of classes can be estimated through the empirical study of the data distribution (as in MSH approach). Repeated iterations of the algorithms starting from different initial point can provide many "possible" partitions which can then be used to formulate a distribution of partitions for inferring the most likely partition. Thus, the stabilization approach which is presented below, shares concepts with data resampling in the generation of population statistics. This primary assumption posed the logical goal that the distribution of iterated partitions will be representative of the "optimal" or "desirable" clustering result. In this thesis we propose a novel algorithmic approach for extracting adequate information for the accurate classification of cervical cancer AW samples by self-organizing the dataset via the combination of k-means clustering and data resampling (bootstrapping) [9]. In particular, we examine the stability of k-means clustering, which aims to derive stable and representative class centers through iterations in the initialization process.

- ***Aim1- Stability Assessment through random resampling and initialization***

This thesis recommends a novel technique for self-organizing data, without any prior knowledge on their statistical distribution, fusing efficient strategies from clustering and resampling. The proposed methodology aims at searching for hidden characteristics within the processed dataset and revealing additional data structures or subclasses that can be utilized for identifying irregular groups that are of particular importance in disease modeling. The performance evaluation of the presented algorithm to biomedical data from cervical cancer is tested and analyzed on sample vectors representing the temporal response of tissue areas obtained through multispectral imaging. The results of this study show that stratified, repeated applications of simple clustering schemes can effectively organize large data, giving rise to the application of the proposed method for tissue giving rise to the application of the proposed method for tissue classification for enabling accurate and early disease diagnosis.

In order to seek and alleviate initialization problems and stabilize the partitions of clustering, such as k-means scheme, we use the Euclidean distance and a fixed number of clusters. Towards this direction and using the smaller training set of 497 samples, we examine

the iterative generation of multiple class centers, under the assumption that the random reshuffling of data generates centroid groups whose topography essentially reveals the class-center locations. Thus, random initializations are expected to produce slightly deviated centroids, as well as outlier centroids. The repetition of many such processes is expected to group meaningful centroids without significant influence from the outliers. Thus the subsequent organization of centroids into hyper-classes that determines and reveals the final set of centers. We evaluate the efficiency of this approach both qualitatively (based on expert's opinion) and quantitatively by constructing the confusion matrix on the labeled 497 samples of the training set. Through the implement analysis scenarios, we examine:

- How well the seven reference curves, even though obtained from clinical experience, can represent the set of 497 samples.

- How well the labeled data samples cluster together as representative curves of different cancer stages.

- How can the data samples be used for self-organization of classes, without any prior information on their state (or labeling) in a completely unsupervised fashion.

- The self-organizing approach follows the MSH approach not for the data samples, but for the centers of "typical" k-means classes produced through iterative re-evaluation of clustering from different initializations. In this form, the cost of MSH is significantly reduced. Here we examine the efficiency of MSH in recovering representative class centers for fixed number of classes (seven classes inspired by the clinical, reference curves), as well as in specifying the number of classes using class compactness metrics, such as the silhouette metric.

### ***Experimental data that are used in the proposed approach***

This proposed methodology is tested on training dataset of 497 reference samples experimentally labeled by a medical expert based on biopsy. Through recursive repetitions of k-means, we produce a large number of class centers, which are then recognized into tight groups through the MSH approach. This last step specifies the appropriate number and the centers of classes. Finally based on these centers all sample curves are organized in classes and the confusion matrix (using the known labels) is used as a means of evaluating the efficiency of the recovered class centers in representing the classes of interest.

As mentioned before, two data sets are available. The first data set contains 497 samples in the form of a (497, 29) matrix, for which we also have information regarding the cellular state of the cancer at this point through actual biopsy (71 instances for each of the seven cancer states). In addition, we are also given a set of 7 reference curves in the form of a (7, 29) matrix reflecting representative responses of the seven clinical states of particular interest. These seven curves have been specified through clinical validation of representative biopsies and reflect the

clinically accepted typical responses of the 7 distinct cancer types. These ground- truth waveforms are illustrated in Figure 7, revealing the differentiation in both shape and amplitude of the characteristic tissue identity curves.

## ➢ *Proposed algorithmic framework*

After the presentation of the data acquisition procedure along with the basic theoretical aspects beyond clustering, the proposed algorithmic framework for cervical tissue evaluation is established. In this thesis a novel, fused self-organization scheme is proposed, in a statistically significant manner based on different groups originated from the dataset after multiple runs of k-means clustering. Ideally, we force k-mean to produce similar statistical populations of class centers, with the generated classes being slightly differentiate in every run due to different initialization. With the view to produce the final class centers, we apply the MSH approach to automatically re-arrange the multiple centroids produced previously. These processes are tested on the IBSL sample, constituting the training stage of our algorithmic framework. In scenario 1, we examine progressively four different cases, enriching the proposed data self-organization scheme with information and evaluating its performance in order to result in the optimal algorithmic setup.

### *I. Case 1: Compatibility of clinical trends with measures responses*

The seven reference curves initially shown at Figure 7 are also presented in the left plot of Figure 10. These curves, which are available in the data, reflect the clinical trends for deciding on the stage of cancer. In this part, which has also been addressed in scenario 1 where it is used as, we explore the ability of these curves to actually capture the distribution characteristics of curves measured from each particular stage. In particular, we use these given reference curves as cluster centers and apply the Euclidean distance metric for the clustering of 497 curves treated as unlabeled data. In the sequel, we use the labels of these curves in order to produce a confusion matrix for the derived partition and evaluate its effectiveness in correctly clustering these measurements.

***Methodology of Case 1***

**Step 1:** Utilize the 7 reference clinical curves as a cluster centers.

**Step 2:** Apply the Euclidean distance metric for the clustering of 497 curves treated as unlabeled data.

**Step 2:** Use the labels of these curves in order to produce a confusion matrix for the derived partition.

### *II. Case 2: Compatibility of biopsy-validated classes with labeled responses*

The 497 labeled curves inherently define data clusters, which can be represented by their class centers. We use the mean of curves within each class in order to derive the corresponding class center. Then, we explore the ability of these centers to actually reflect the distribution of classes using the previous computational framework. The 7 labeled centroids are depicted in the right plot of Figure 10.

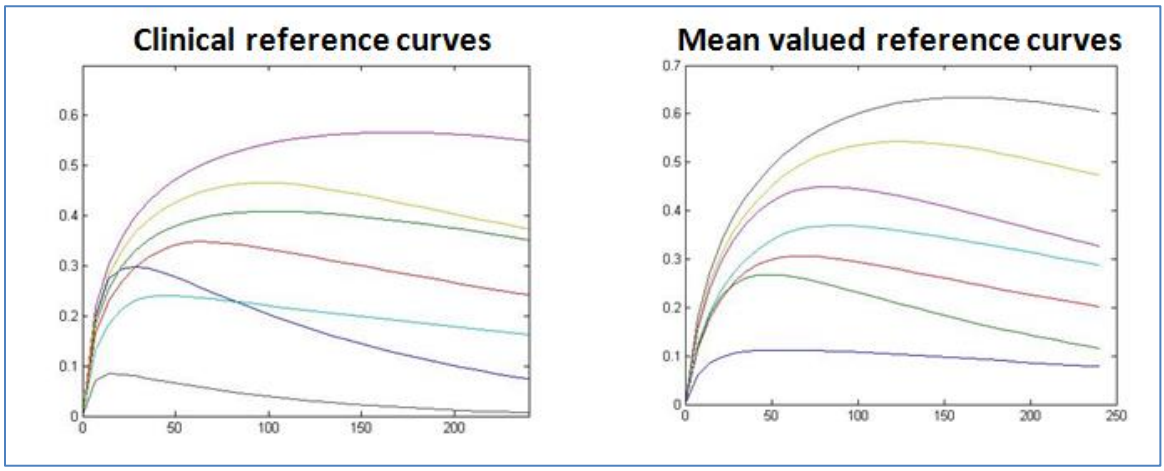


Figure 10. Qualitative illustration of the 7 clinical reference curves, reflecting the cancer stages in the left plot and the labeled centroids regarding the mean of curves of each of labeled classes.

***Methodology of Case 2***

**Step 1:** Use the mean of curves within each class in order to derive the corresponding class center.

**Step 2:** Apply the Euclidean distance metric for the clustering of 497 curves treated as unlabeled data.

**Step 3:** Use the labels of these curves in order to produce a confusion matrix for the derived partition

III. *Case 3: Effectiveness of self-organization of data in a specific number of classes*

In this case, we explore the possibilities for self-organization of data into seven classes, using distance-based clustering. In particular, we utilize the simple k-means approach to organize the available data into seven classes, starting from one initial point. Since this initialization is driving the final partition, we repeat the approach many times sampling on the initialization points. Thus, we generate a number of partitions with associated cluster centers, with reflect a representative distribution of classes or class centers. At this stage we use the MSH approach applied only on the set of class centers as to organize and retrieve seven classes of centers reflecting the seven stages of interest. The mean of each cluster of centers defines the overall

class center. The final set of class centers is evaluated by clustering the 497 samples and computing the confusion matrix of this partition.

***Methodology of Case 3***

**Step 1:** Generate a number of partitions via multiple runs of k-means clustering on the dataset with 7 classes.

**Step 2:** Use the MSH approach applied only on the set of class centers so as to retrieve seven classes of centers.

**Step 3:** Utilize the labels of these curves in order to produce a confusion matrix for the derived partition.

IV. *Case 4: Self-organization of data with self-evaluation of the number of classes*

The final case examines the stabilization potential of the re-sampling initialization scheme for fully automated data organization, without any prior knowledge. Our proposed scheme proceeds similar to the previous case via repeated application of the k-means to produce a population of class centers. In particular, we process the set of 497 samples in a bootstrap mode repeatedly running k-means algorithm for 50 times and a fixed number of classes k=10 so as to generate and organize a population of 500 cluster centroids, adopting the Squared Euclidean distance metric, which is proved to recognize overall amplitude similarities. Then the centroids dataset is refined via a MSH clustering scheme without any knowledge on the number of classes, which produces the final clustering centers and the number of dominant classes that optimize the Silhouette criterion [10]. The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the $ith$ point, $s(i)$, is define in equation:

$$s(i) = \frac{A_{1,i} - A_{2,i}}{\max\left\{A_{1,i}, A_{2,i}\right\}} \qquad (4.1)$$

where $A_1$, is the average distance from the $ith$ point cluster to other points in the same cluster as $i$ and $A_2$ is the minimum average distance from the $ith$ point to points in a different cluster, minimized over clusters. As a validation stage, we utilize the labels of the seven given reference IBSL curves in order to produce a confusion matrix for the derived partition and evaluate qualitatively the effectiveness of the algorithmic scheme in correctly clustering the input tissue status measurements.

As a validation stage, the final set of class centers is also tested through the confusion matrix, but since the number of labels might be different from the number of classes, the final distribution of classes is qualitatively evaluated through the visual inspection of classes and

quantitatively supported by compactness measures. The methodology steps are described below and the algorithmic framework to implement the above case is illustrated at Figure 11.

***Methodology for Case 4***

**Step 1:** Repeatedly apply k-means algorithm to produce a population of class centers.

**Step 2:** Apply the MSH scheme to this population without any knowledge on the number of classes, using every time Silhouette metric to evaluate automatically.

**Step 3:** Firstly, use labels of these curves in order to produce a confusion matrix for the derived partition.

Secondly, qualitatively evaluate final distribution through the visual inspection of classes.

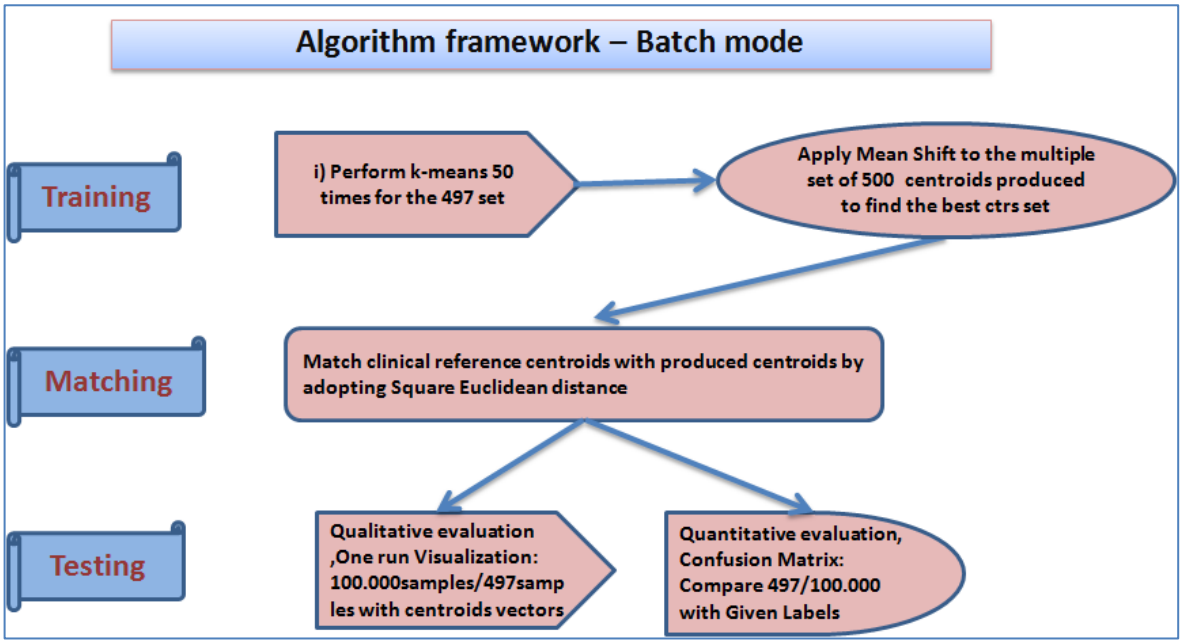


Figure 11. Proposed algorithmic framework implementing the fusion of clustering and data resampling approach.

### 4.2.2 Scenario 2: Clustering for Self-organization of Dynamic Imaging Data: Developing a Recursive-Mode K-Means

The term of "center-based clustering" is considered to a wide category of clustering methodologies (including k-means and Gaussian-Expectation Maximization), implying the utilization of a number of "centroids" to represent and partition the input data. Such algorithms begin with a guess about the solution and then refine the positions of centers until they converge to a local optimum, which however can be far from the global minimum. This is a primary problem of data clustering, related to the sensitivity to initialization of the seed cluster centers. Towards this direction, many "wrapper" techniques transform the input or the output of the clustering procedure and/or perform multiple runs of the algorithm [7, 8]. In this study we focus on the improvement of the classic k-means approach, by exploring the potential to directly make it less sensitive to initializations and give better solutions, refining cluster centers in a recursive mode, averaging the currently and previously calculated value.

Within the data space, the optimization criterion of the classical k-means approach calls for the minimization within each call of the Euclidean distance from the class center. This formulation adopts the full information from the data distribution, simulating the conditional class distributions based on the available data. Enhancing this approach and incorporating the prior expert knowledge by means of the prior class distributions, we formulate the criterion conditioned on the distance from the previous centers. This modification using both prior and conditional distributions. Our tested scheme gives same weight to the prior and conditional terms (equal to ½), but the formulation can utilize a variable weight of combining the two terms. In addition, our scheme utilized the Euclidean distance in the two terms, but more robust distance metrics could be tested within a more computationally complex and time-demanding framework.

- ***Aim 2 – Stability Assessment through recursive refinement of cluster centers***

In this study we present a novel algorithmic approach for automated clustering of time-series data in order to extract adequate information and accurately classify cervical cancer AW samples by controlling the initialization and cluster center update procedure. In particular, we examine the robustness and accuracy of k-means clustering in an iterative mode, which aims to improve data partition, considering current and previous information, instead of completely updating the centers as the classical k-means implementation. In addition, we combine initial knowledge regarding the class centers with the data distribution model and examine the effectiveness of a relaxation scheme in preserving the structural differences of the dominant classes/clusters.

## ➢ *Proposed algorithmic framework*

We consider the organization of data based on unsupervised computation of the population centroids, adopting a k-means based on clustering scheme modified in terms of both initializations and cluster center updates/refinements after each iteration, enabling the improvement and robustness of the extracted results. The main idea behind k-means lies on the minimization of an objective function, generally chosen as the total distance between all patterns from the corresponding cluster centers. The basic concept of k-means has been described in Chapter 1. The distribution of objects among clusters and the updating of the centroids constitute the two basic steps of the algorithm. K-means clustering approaches, although simple and relatively fast and adaptive suffer from poor initialization, implying that different initialization of centers produces different results. Alleviating these standard errors is a challenging issue in statistical analysis. In this paper, in order to improve the performance of k-means clustering, we apply a novel framework to assign class centroids and calculate the labels of the testing dataset based on a weighted cluster updating scheme among information derived from the previous and current iteration and on targeted initialization originating from expert knowledge on the data. The concept of new Recursive k-means mode is described in the following procedural steps.

***Recursive k-means algorithm***

- ***Step 1****. Force k-means to run from initial seeds producing 7 new centroids*
- ***Step 2.****Match produced centroids with initial centroids in order to find the optimal distance between the two different sets: Measure the distance between the initial centroids and the produced centroids. For each initial centroids find the closest new centroids and calculate the average of them.*
- ***Step 3****. Iteratively repeat the above steps until convergence is achieved: Run k-means by using the new centroids each time, as initial seeds. In addition, continue with the matching procedure, updating the new centroids. When the produced centroids don't move any more, keep this set of centroids as the final set.*
- ***Step 4.*** *Validation of the iterative scheme: Utilize the Euclidean distance metric between the group of 497 curves, treated as unlabeled data and the final centroids, with the aim of classifying into groups the 497 clinical cases.*
- ***Step 5****. Evaluation of the results: Compare the derived partition and the known labels of these curves through confusion matrix production, with the view to evaluate the results.*

The proposed algorithmic framework for cervical tissue evaluation is illustrated in Figure12. We propose a novel, self-organization scheme based on the initialization and

centroids recalculation procedure of the classic k-means approach. More specifically, we force k-means to begin from seeds imposed by the medical expert knowledge accompanying our clinical dataset and gradually improve the calculated cluster centers by averaging current and previous class-centroids, so as to avoid updates that may deviate considerably from the candidate optimal solution. Besides the iterative training, the proposed approach needs to match previous and new centroids in order to update them appropriately. These processes are tested on the IBSL sample dataset and evaluated quantitatively and qualitatively through the construction of confusion matrices.

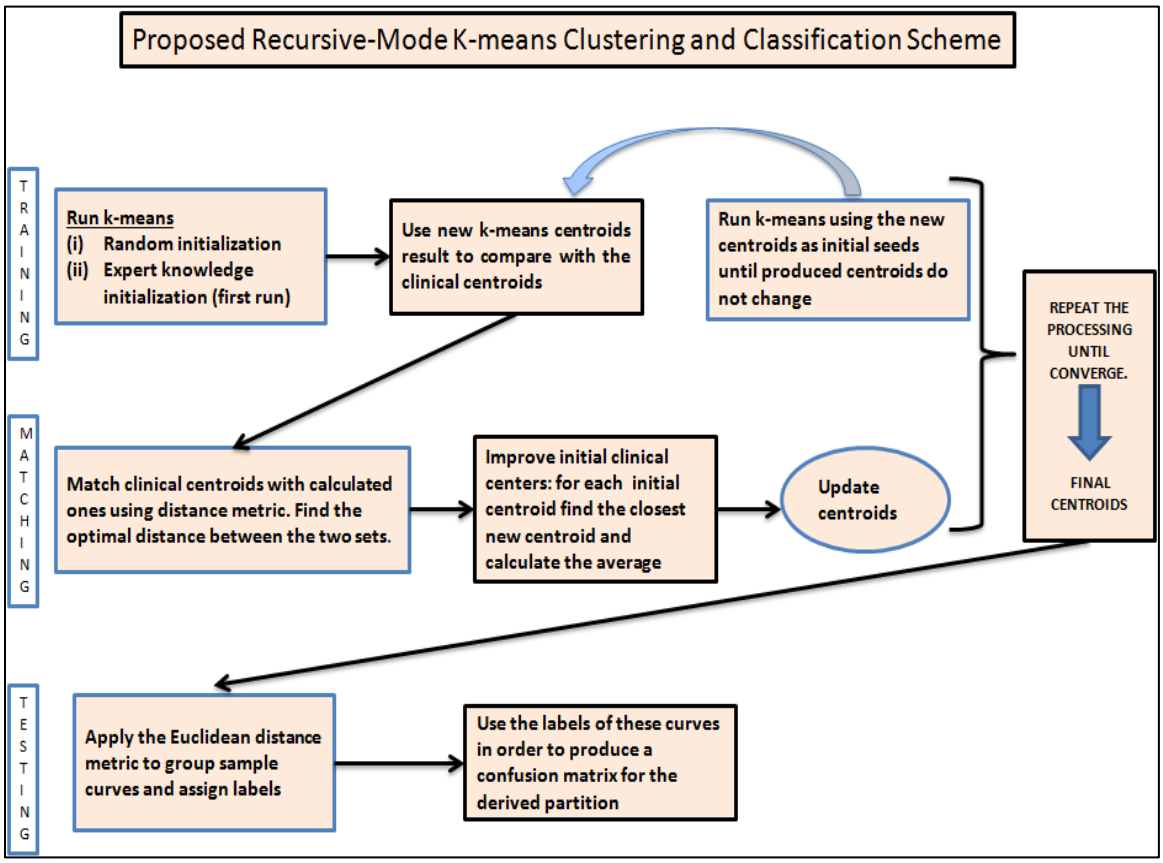

Figure12. Algorithmic scheme of the proposed recursive-mode k-means clustering methodology.

Two different scenarios are presented, based on expert knowledge and random initialization of the clustering procedure and evaluating its performance in order to examine the impact of starting from different seed point to the final cluster assignment and sample classification process. In order to produce comparative results, the proposed methodology is tested against the classic k-means approach performing again the two basic testing cases (random or defined initial points).

### 4.2.3 Scenario 3: Introduce a New Efficient Distance Metric

- ***Aim 3: Exploration of Distance metric and Class Formation Strategy***

In attempt to improve clustering and the partition of the feature space, we study the effects of the distance metric. Euclidean distance can deal with magnitude differences while the Squared Cosine metric can identify shape similarities in structured curves. The application of each metric results in different partitions, determined by the metric properties; one partition groups together curves of similar magnitude (of different shapes) and the other merges all data samples (curves) of similar shape despite their magnitude. Based on the dataset, one can easily identify two characteristic curve shapes, one of rapid increase and decrease and the other with smoother increase and much shower decrease.Thus, in this scenario we explore two different cases in order to succeed the optimal results.

Firstly, we suggest an alternative consideration of class formation that proceeds with the sequential improvement of classes based on different characteristics, i.e. first organize similar magnitude classes based on the Euclidean distance and then split these classes on the basis of shape and the squared-cosine metric.We evaluate the new approach again through the confusion matrix on the labeled 497 samples of the training set and qualitatively the results on the larger 100000 dataset. We evaluate how well the produced class centers represent the labeled data samples by visualizing them.

Secondly, based on the need to derive a distance metric sensitive to both amplitude and shape differences of time-series, we proposed a combination of the two individual metrics (Euclidean and Cosine) with first given to shape considerations.

- ***Proposed algorithmic framework***

1 ***Efficient Assessment of sequential clustering***

In this scenario, we explore the possibilities of a combined clustering approach through the sequential application of distance metrics. More specifically, we first target the organization of data using the Euclidean distance and then the refinement of each class using the squared-cosine distance. The first partition results into seven initial classes, while the second stage produces fourteen classes, which fulfill both distance and shape-based criteria in clustering. In particular, we first classify the testing set of 100000 curves given the 7 reference centroid vectors derived via clinical validation into seven groups that provide the minimum Squared Euclidean Distance value.

At a second stage, we investigate the potential of revealing hidden sub-classes of considerable structural similarity, splitting each of the seven groups derived into two smaller ones performing a k-means clustering scheme with k=2 and the Squared Cosine metric as the objective function.

This sequential clustering strategy exploits both amplitude and shape similarities within the data population. In order to avoid any class over-representation through the splitting process, we also follow a merging process for classes whose centers fulfill certain similarity criteria in magnitude and shape terms. The proposed sequential scheme is qualitatively evaluated examining the shape and size characteristics of the samples that form each of the finally

extracted clusters. This sequential scheme generally achieves good within class characteristics, but results in many classes with unclear interclass relations. It us our conclusion that this form of approaches needs both splitting and merging provisions for classes, resulting in cumbersome and time-consuming procedures. The next section considers a combined (unique) distance instead of the sequential application of multiple distance metrics.

## 2 Combined distance metric

Different approaches of k-means clustering focus on changing the basic objective function to measure the dissimilarity between an object and a certain class, leading to different representation of clusters. The basic k-means method utilizes the $L_2$ norm distance to express similarity and normally considers the centers of each cluster as the mean of data it preserves. Although Euclidean based distance metrics perform well in identifying differences in amplitude within the overall data range, the results they produce when comparing similarity sized but unequally shaped data can be misleading. On the other hand, Squared Euclidean Cosine distance metric which also was selected, recognizing shape similarities although it losses the magnitude. In this scenario 3, we explore the possibilities of a new combined distance approach based on the Euclidean and the Cosine distance metric. The proposed combined distance metric is given as:

$$D(A1, A2) = \underbrace{\left(1 - \frac{\left(\sum_i A_{1,i} A_{2,i}\right)^2}{\sum_i A_{1,i}{}^2 \sum_i A_{2,i}{}^2}\right)}_{(a)} + \underbrace{g * \left(\sum_{i=0}^{n} \left(A_{1,i} - A_{2,i}\right)^2\right)}_{(b)} \quad (4.2)$$

where the former part (a) reflects the Squared Cosine distance, for $A_1, A_2$ representing the two sample curves to be compared, and $i$ notifies the time points of curve measurements. The latter part (b) of the combined matric introduces the Euclidean distance multiplied by a factor g. From experimental tests, we propose to utilize value of g computed as follows:

$$g = \frac{1}{n_i} \sum_{i=1}^{n_i} \max_j (c_{i,j}) \quad (4.3)$$

where $c_{i,j}$ denotes the value of the $ith$ centroid curve at the $jth$ time stamp, whereas $n_i$ indicates the number of centroids. Thus, the scale factor g is determined from the average maxima of the centroids curves.

In experimental validation stage, we explore the potential of the new combined distance and test how well it fulfils both amplitude and shape-based requirements on the set of 497 samples. First we test the individual Euclidean and Cosine metrics on the clustering of this population, using already derived centers from the previous stage of development as well as the clinically approved centroids. In particular, we cluster the set of 497 curves based on their

minimum distance from the set centroids. This test illustrates the problem of either distance metric and highlights the need of a combined metric. In the sequel, we utilize the proposed distance metric within the stable clustering scheme presented in scenario 1 and Case 4. Finally, we formulate the confusion matrix for this strategy, using again the true labels of the IBSL curves.

### 4.2.4 Scenario 4: Bootstrap Clustering Approaches for Self-Organization of Big Data

Bootstrapping is general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand. The aimed statistic in our application is the centroid (or mean) vector, which follows a multimodal distribution depending on the number of classes reflected in the dataset. Thus, the process of randomized resampling aims at deriving a large number of potential class centroids, building in this way an empirical multimodal distribution for the centroid statistic. We apply this scheme in two concretevariations, which also are depicted in Figure 13, notice that the use of the proposed combined distance is embedded in all processing steps:

1) Instead of resampling the data, we resample the presentation sequence of this data as to build the sample distribution of the statistic of interest, which is actually the mean vector (centroid) of classes. This could also be interpreted as a permutation approach, where the sequence of samples is randomized instead of a characteristic class label.
2) Resample data from a large set (big data) in the form of multiple subsets that reflect subparts of the entire data space, in order to piece wisely (or sequentially) construct the sampling distribution of the test statistic in the original space.

Subsequently, this distribution of the summary statistic is discretized into a number of concrete vectors representing the multimodal peaks, though a second clustering approach applied on the metadata of potential class centroids.

In this study [11] we explored the potential of repeated iterations in k-means clustering to derive stable partitions of fixed number of classes. The basic aim of the scenario 1 focused on forcing k-means algorithm to repetitively produce multiple sets of cluster centers reflecting the ensemble of possible classes generated from the same population but with different initialization. At a second stage, the cluster centers are groups via the MSH approach, which aims at automatically re-arranging the multiple centroids produced within the initial stage. In this work, we expand this framework in order to accommodate not only stability but also generalization aspects to the autonomous organization and clustering of large datasets. In essence, we formulate multiple clustering operations on smaller bootstrap subsets with the aim to cover the entire data and partition spaces with multiple randomized subsets, which can be partitioned more efficiently and, as a whole, accommodate the entire dataset of significant size. Along these analysis issues, we also apply in this process the alternative distance metric which we introduced in previous scenario 3 and evaluation schemes in order to enforce both amplitude and shape similarities within each class formation.

This concept is tested on cervical cancer staging with the corresponding IBSL curves without any prior information on the data nature and properties, attempting to extract the hidden information through resampling methodologies. Considering the extracted results, the efficiency, robustness and flexibility of the proposed exploratory data analysis framework is evaluated.

Partitioning the data space via a generic clustering algorithm is determined by the adopted distance metric and the selection of class centers with represent the class characteristics, including the number of dominant clusters, the cluster size and shape. All these issues are affected by initialization. Furthermore, the generalization of partitions when increasing the dataset with new sample is an issue of particular importance in clinical applications and categorization of new cases, which has not been treated in due care. As already mentioned, we develop a self-organization framework to deal with these issues. The development proceeds addressing the following key direction, which is validated on the available data from cervical cancer response curves. This successive consideration along with the implementation step is described below.

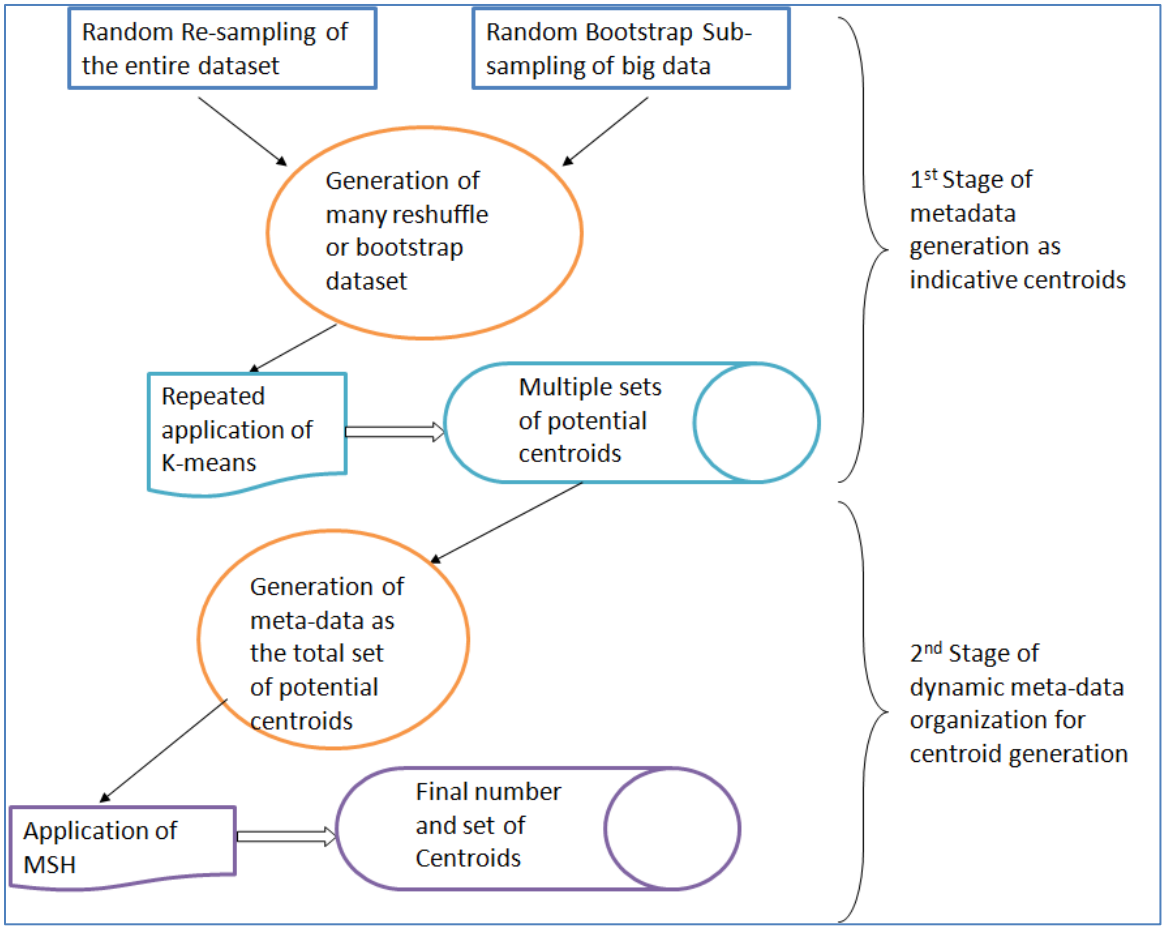


Figure 13. Flow diagram of algorithmic developments; Stage 1 employs either random initializations (for stability) or random bootstrap subsets (for generalization) and generates meta-data as the set of potential centroids; Stage 2 operates on meta-data through the MSH scheme and organizes the potential groups into more robust centroids.

- ***Aim 4: Generalization Assessment through bootstrap datasets of small size***

At this stage, we explore generalization aspects of the bootstrap clustering scheme on a large number of 100.000 IBSL curves. We attempt to cover the data space with multiple,

randomly selected, smaller subsets and re-combine at a higher-organization level. In the process of clustering of each subset in the iterative bootstrapping scheme, we exploit the previously developed stabilizing process in scenario 1. We also utilize the combined distance developed in the lines of scenario 3, which absorbs both Squared Cosine and Euclidean metrics in order to perceive information based on both size and shape characteristics of the time-series or IBSL curves. Furthermore we test if the derive clustered centers can produce efficient grouping and unmixing of the large dataset. Towards the evaluation of this complete self-organization framework we exploit both qualitative and quantitative means. First we compare the class distributions and the class centers extracted from the application of the proposed algorithmic scheme to the labeled dataset of 497 samples and the totally unknown large dataset of 100.000 samples. Then, we formulate the confusion matrices for the 497 labeled and clinically validated samples using the class centers derived from the strategy using bootstrap subsets and compare them with the ones extracted by processing only labeled curves, along the development of scenario 1 and 3 above.

- ***Proposed algorithmic framework***

This case examines the potential for fully-automated self-organization of data, without any prior knowledge on its origin and nature, searching for hidden structured groups within a large dataset, with self-evaluation of the number of classes and their class centers. In particular, this step explores the generalization potential of bootstrapping by forming small datasets and using their structure to organize hyper-clusters that can generalize the large dataset. The proposed framework of self-organization enables the generation of subclasses that are not influenced from prior clinical knowledge, the comparison of data-driven clusters with clinically relevant disease, as well as the derivation of new sub-classes originating from the data, which deserve further clinical consideration.

Our proposed initially proceeds via repeated application of the k-means algorithm with the combined distance metric, in order to produce a population of potential class centroids. This step performed on 250 bootstrap sets, each deriving a fixed number of k=10 centroids in each run. Then, the bootstrap set of 2500 centroid vectors is clustered via an MSH scheme, which automatically evaluates the number of dominant classes through the minimization of the Silhouette metric and produces a more robust and representative grouping of meta-data (i.e. centroids) at a higher-abstraction level. The final set of class centers is tested through the confusion matrix, but since the number of labels might be different from the number of classes, the final distribution of classes is qualitatively evaluated through the visual inspection of classes and quantitatively supported compactness measures.

The above stages of development highlight different aspects and provide solution to varius problems of exploratory clustering algorithms. Altogether they form a framework of operation that can be applied to big data analysis towards deriving classes directly from the data. The core idea is the generation of meta-data that can be seen as the set of all potential class centroids and forms an attempt to extract information from the original data at a higher-abstraction level. This set is effectively processed by the MSH scheme, which operated on the meta-data

distribution space. Nevertheless, the initial data resampling is used, whereas bootstrap techniques are employed to assess generalization issues.

# References


1) William P. Soutter,Emmanuel Diakomanolis, Deirdre Lyons, "Dynamic Spectral Imaging: Improving Colposcopy", Clin Cancer Res, Vol. 15, No. 5, pp. 1814-1820, 2009.
2) C. J. Balas et al. “In vivo detection and staging of epithelial dysplasias and malignancies based on the quantitative assessment of acetic acid – tissue interaction kinetics”, J. Photochem. Photobiol. B-Biology, Vol. 53, No. 1–3, pp 153–157, 1999.
3) A.K Jain, M.N Murty and P.J Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol 31, No.3, pp. 264-323, September 1999
4) A.K Jain, “Data Clustering: 50 Years Beyond K-means”, Pattern Recognition Letters, Vol 31, pp 651:660,2010.
5) Argyris Kalogeratos and Aristidis Likas, “Dip-means: an incremental clustering method for estimating the number of clusters”, Advances in Neural Information Processing Systems, Vol. 25, pp. 2402-2410,2012.
6) Donglin Niu, Jennifer G. Dy and Michael I. Jordan, “Iterative Discovery of Multiple Alternative Clustering Views”, Vol 6, Issue 7, pp. 1340-1353, doi 10.1109/TPAMI.2013.180,2013.
7) A. Likas, N Vlassis, and J. Verbeek., “The global k-means clustering algorithm.”, Technical report, Computer Science Institute, University of Amsterdam, The Netherlands, IAS-UVA-01-02, February 2001
8) J. Pena, J. Lozano, and P. Larra naga, “An empirical comparison of four initialization methods for the k-means algorithm”, Pattern recognition letters, Vol. 20, pp. 1027-1040, 1999
9) T. Warren Liao, “Clustering of time series data-a survey”, Pattern Recognition, Vol. 38, pp. 1857-1874,2005.
10) C. A. Field and A. H. Welsh, “Bootstrapping clustered data”, Journal of the Royal Statistical Society, Series B, Vol. 69, Part 1,3, pp 369-390, 2007.
11) I. Vourlaki, G. Livanos, M. Zervakis, C. Balas, G. Giakos, “Spectral Data Self-organization Based on Bootstrapping and Clustering Approaches”, Accepted for presentation in IEEE Imaging Systems and Techniques (IST) Conference, Macau, China, 16-18 September, 2015.

# 5. Results

## 5.1 Scenario 1: Data Self-Organization through Resampling and Clustering Approaches without Prior Knowledge Results

The results in this scenario 1 illustrate the qualitative and quantitative organization of data based on the algorithmic implementation which is described in previous section for each case. In particular, the evaluation of the proposed methodology is based on the construction of the confusion matrix of estimated labels against the ground-truth ones, assigned through tissue biopsy and characterization. The analytical description of confusion matrix has been addressed in the section of results in previous Chapter 4.

### ✓ Case 1: Compatibility of clinical trends with measured responses

As mentioned before, we examined the compatibility of clinical trends with measured responses. We highlight the ability of the stabilized clustering scheme to correctly capture these clinical trend by just the available data curves. The seven reference curves form the clinically approved responses of the different stages of the pathology and reflect the clinical trend for deciding on the stage of cancer. The results are depicted at Figure 8 & Figure 9.

The 7 clinical reference curves are also applied to the testing, big dataset of 100000 samples, by performing Euclidean Distance, constructing the new confusion matrix based on the reference labels given. It is important to mention that this dataset along with the reference classification originate from simulation procedures based on models validated on clinical data. The confusion matrix, which is illustrated at Figure 14, clearly reveals the difficulty in properly discriminating the estimated classes. The results of Case 1 to this new data set are depicted in Figure 15, where from the representation of the dataset of 100000 curves, it is clear that there is a problem for clinical reference centroids to classify efficiently the big data set.

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
|---|---|---|---|---|---|---|---|---|
| 1 | 1972 | 8972 | 0 | 0 | 0 | 0 | 23154 | 34.098 |
| 2 | 1413 | 17475 | 3 | 0 | 0 | 0 | 0 | 18.891 |
| 3 | 0 | 13405 | 8376 | 0 | 0 | 0 | 0 | 21.781 |
| 4 | 0 | 0 | 9294 | 7779 | 0 | 0 | 0 | 17.073 |
| 5 | 0 | 0 | 0 | 2158 | 1578 | 0 | 0 | 3.736 |
| 6 | 0 | 0 | 0 | 66 | 1429 | 2307 | 0 | 3.802 |
| 7 | 0 | 0 | 0 | 0 | 0 | 619 | 0 | 619 |
| count | 3385 | 39852 | 17673 | 10003 | 3008 | 2926 | 23154 | |

Figure 14. Confusion matrix applying Case 1 to dataset of 100000 simulated classes.

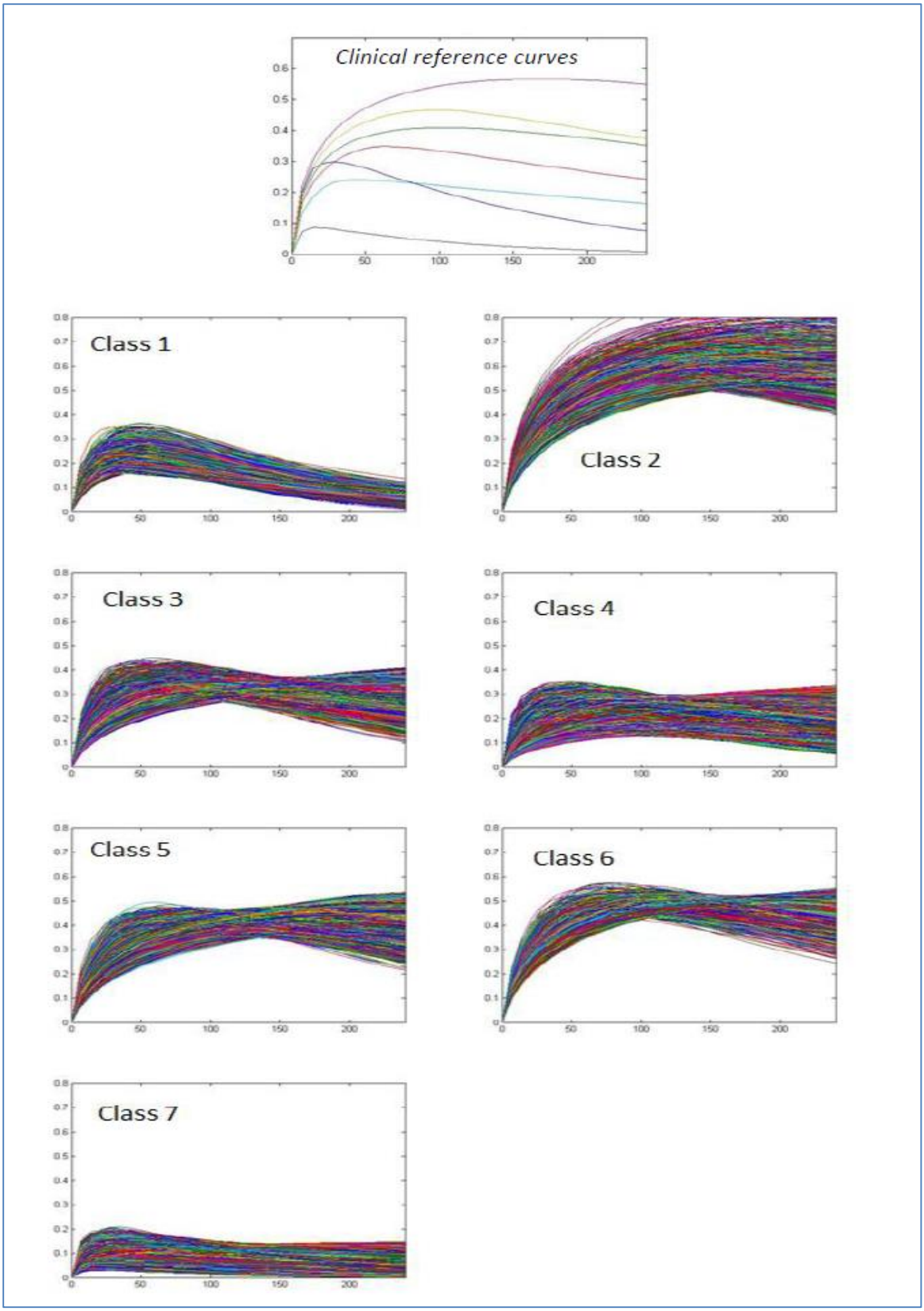


Figure 15. Qualitative illustration, based on the clinical trends to dataset of 100000 simulated classes.

### ✓ Case 2: Compatibility of biopsy-validated classes with labeled responses

The 497 labeled curves inherently define data clusters, which can be represented by their class centers. In Case 2, we first target the organization of 497 data set using the Euclidean distance based on the centroids obtained as class means of the given 497 labeled curves referred to as labeled centroids. Particularly, we use the mean of curves within each class in order to derive the corresponding class center. Then, we explore the ability of these centers to actually reflect the distribution of classes using the previous computational framework. The improvement in class representation against Figure 8 in Case 1 is depicted in Figure 16.

***Classification Results (derived labels per class)***

| *Reference labels per class* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **class** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **count** |
| **1** | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| **2** | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| **3** | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 71 |
| **4** | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| **5** | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 71 |
| **6** | 0 | 0 | 0 | 0 | 0 | 69 | 2 | 71 |
| **7** | 0 | 0 | 0 | 0 | 0 | 13 | 58 | 71 |
| **count** | 71 | 71 | 71 | 71 | 71 | 82 | 60 | |

Figure 16. Calculated confusion matrix regarding the implementation of Case 2: Apply Euclidean distance to the taraining set for 497 labeled IBSL curves with labeled centroids estimated through the labeled population.

The use of prior information is not always applicable but a way to test data classes compactness. By exploiting the labeling information from the examined population to extract the averaged reference IBSL curves, we manage to reach more accurate tissue characterization. We accept that this case is not appropriate for actual characterization of data, since it utilizes prior information on the data distribution through the exploitation labels, but it is only used to indicate that the data classes are quite compact as to allow for automated clustering attempts. In Figure 17 it is depicted how well the seven labeled centroids classify the 497 data set in seven groups, which approximate preserve their shape.

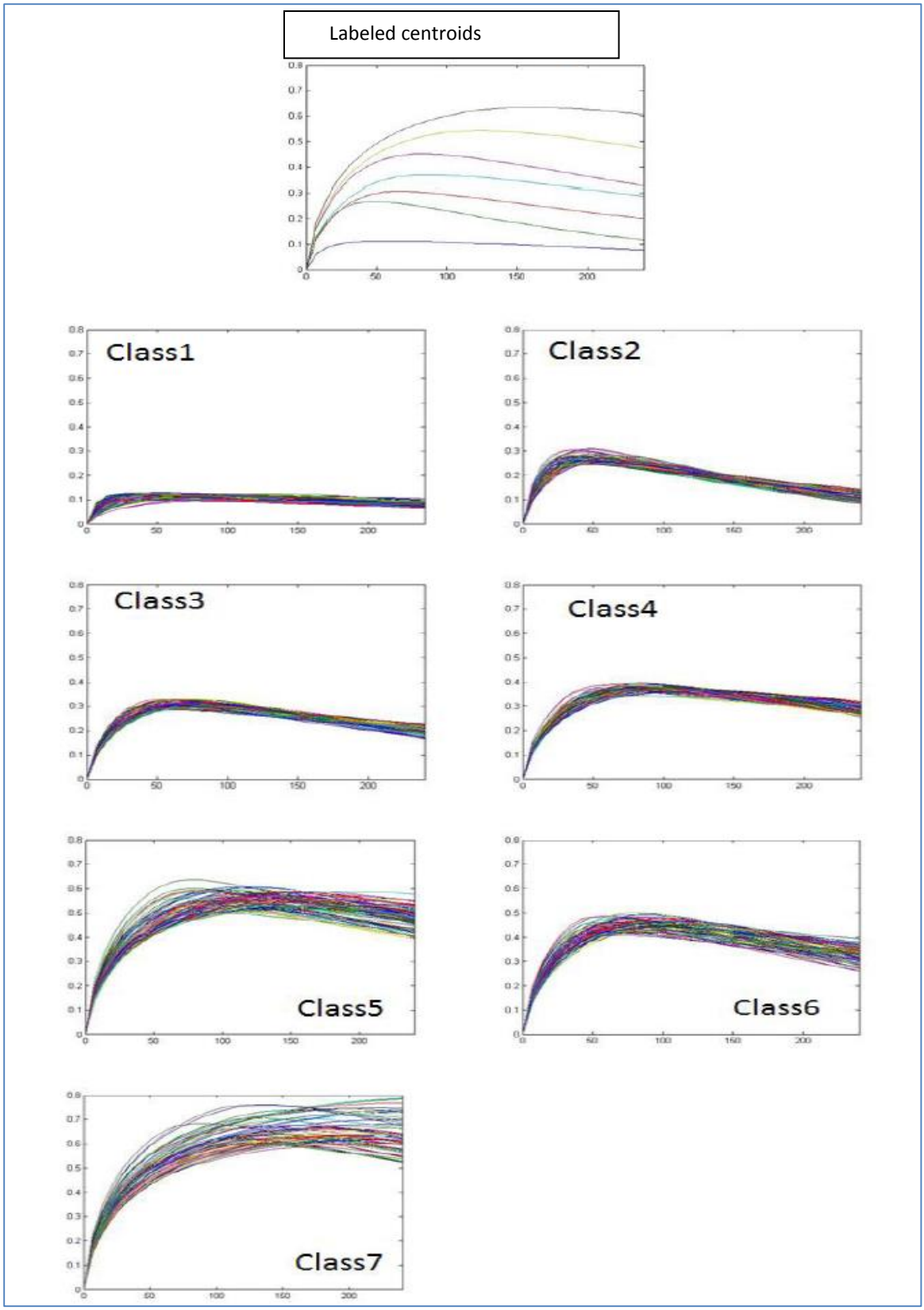


Figure 17. Qualitative representation of 497 IBSL curves regarding the implementation of Case 2 based on 7 labeled responses curves.

In second stage, we classify the testing set of 100000 curves given the 7 labeled centroid vectors derived via the calculation of the mean of curves within each class from training set of 497 curves, into seven groups that provide the minimum Euclidean distance value. We formulate the confusion matrix for the 100000 curves against the ground-truth labels, which is illustrated in Figure 29.

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
|---|---|---|---|---|---|---|---|---|
| 1 | 33660 | 438 | 0 | 0 | 0 | 0 | 0 | 34.098 |
| 2 | 19 | 18847 | 25 | 0 | 0 | 0 | 0 | 18.891 |
| 3 | 0 | 1508 | 20076 | 197 | 0 | 0 | 0 | 21.781 |
| 4 | 0 | 0 | 410 | 15021 | 1633 | 9 | 0 | 17.073 |
| 5 | 0 | 0 | 0 | 0 | 3409 | 327 | 0 | 3.736 |
| 6 | 0 | 0 | 0 | 0 | 0 | 3428 | 374 | 3.802 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 619 | 619 |
| count | 33679 | 20793 | 20511 | 15218 | 5042 | 3764 | 993 | |

Figure 18. Confusion matrix applying Case 2 to the dataset of 100000 simulated classes. The labels are produced via the implementation of Euclidean distance based on the 7 labeled curves vectors.

The results of applying Case 2 to the new large dataset of 100000 IBSL curves, are depicted in Figure 19, demonstrating that the labeled centroids producing significant mixed classes.

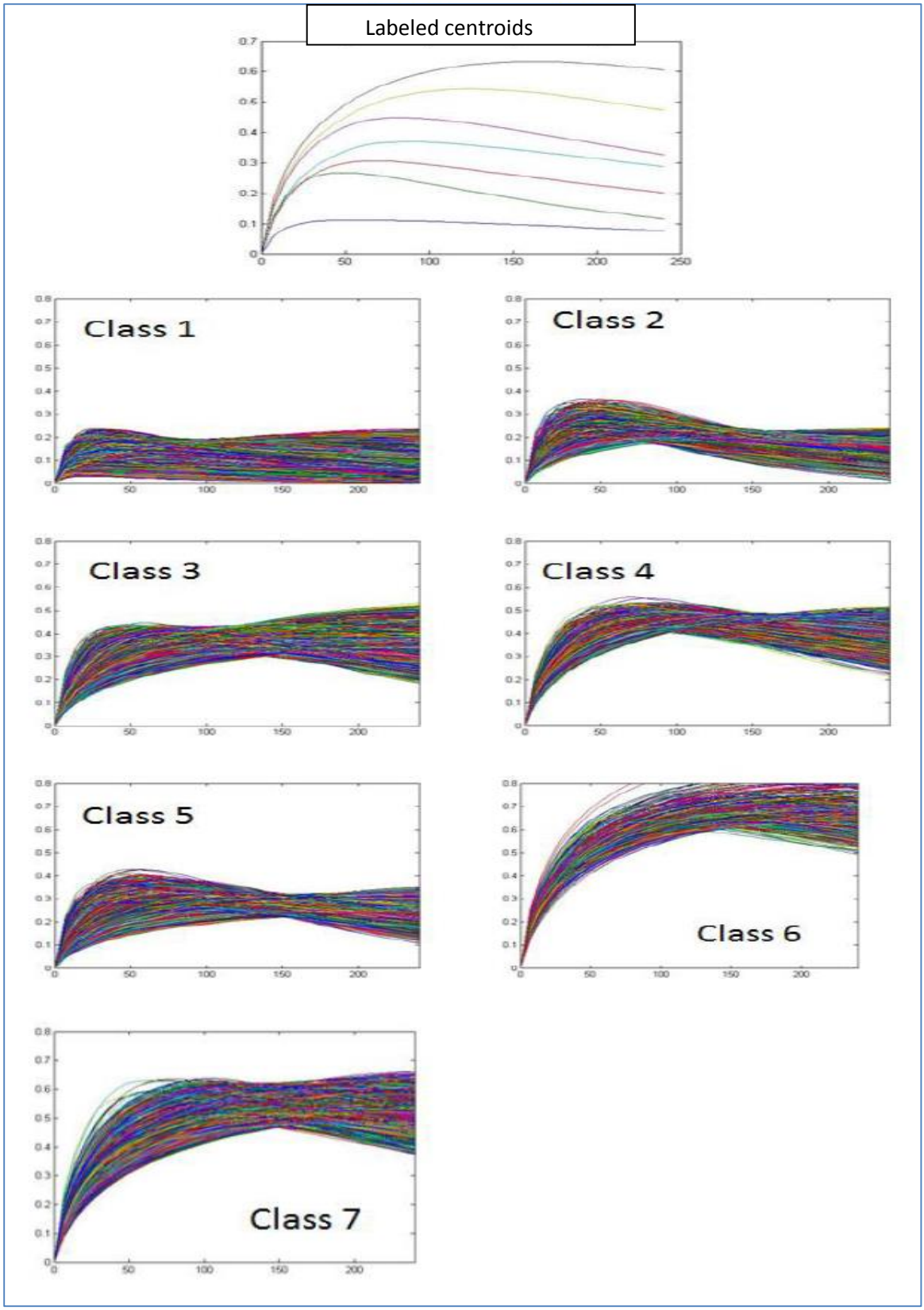


Figure 19 Qualitative illustration of the big data set of 100000 curves based on the implementation of Case 2.

### ✓ Case 3: Effectiveness of self-organization of data in a specific number of classes

In this case, we present a more realistic form the automated organization of data, with the resampling of the training set along with the refinement of estimated centroid curves via MSH clustering of a population of class centers produced by k-mean on a fixed number of 7 classes. The mean-valued labeled centroids with respect to the produced ones based on this approach, are depicted in Figure 20. The corresponding confusion matrix is illustrated in Figure 21.

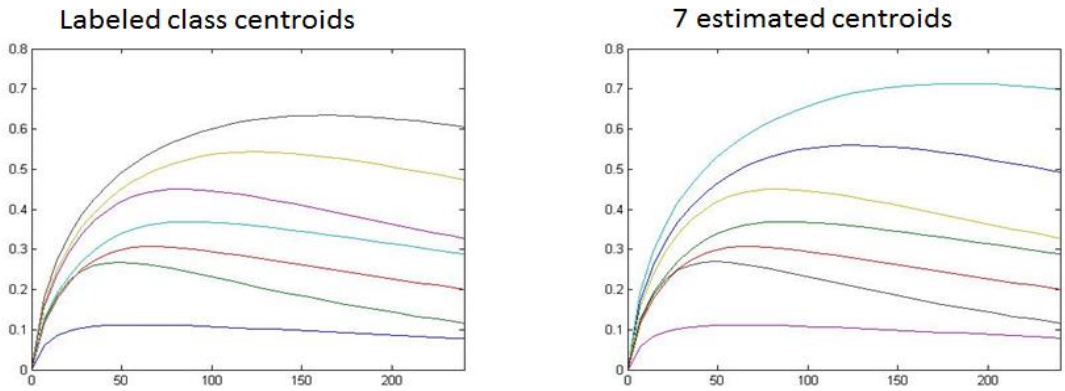


Figure 20. Labeled (left) and derived ones (right) regarding the implementation of Case 3.

***Classification Results (derived labels per class)***

| *Reference labels per class* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **class** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **count** |
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 71 |
| 7 | 42 | 0 | 0 | 0 | 0 | 0 | 29 | 71 |
| **count** | 113 | 71 | 71 | 71 | 71 | 71 | 29 | |

Figure 21. Quantitative illustration of results via confusion matrix along with the estimated IBSL curves regarding the implementation of Case 3.

Final, the ability of the approach to correctly derived all classes of curve shapes except for Class 7, it is represented in Figure 22.

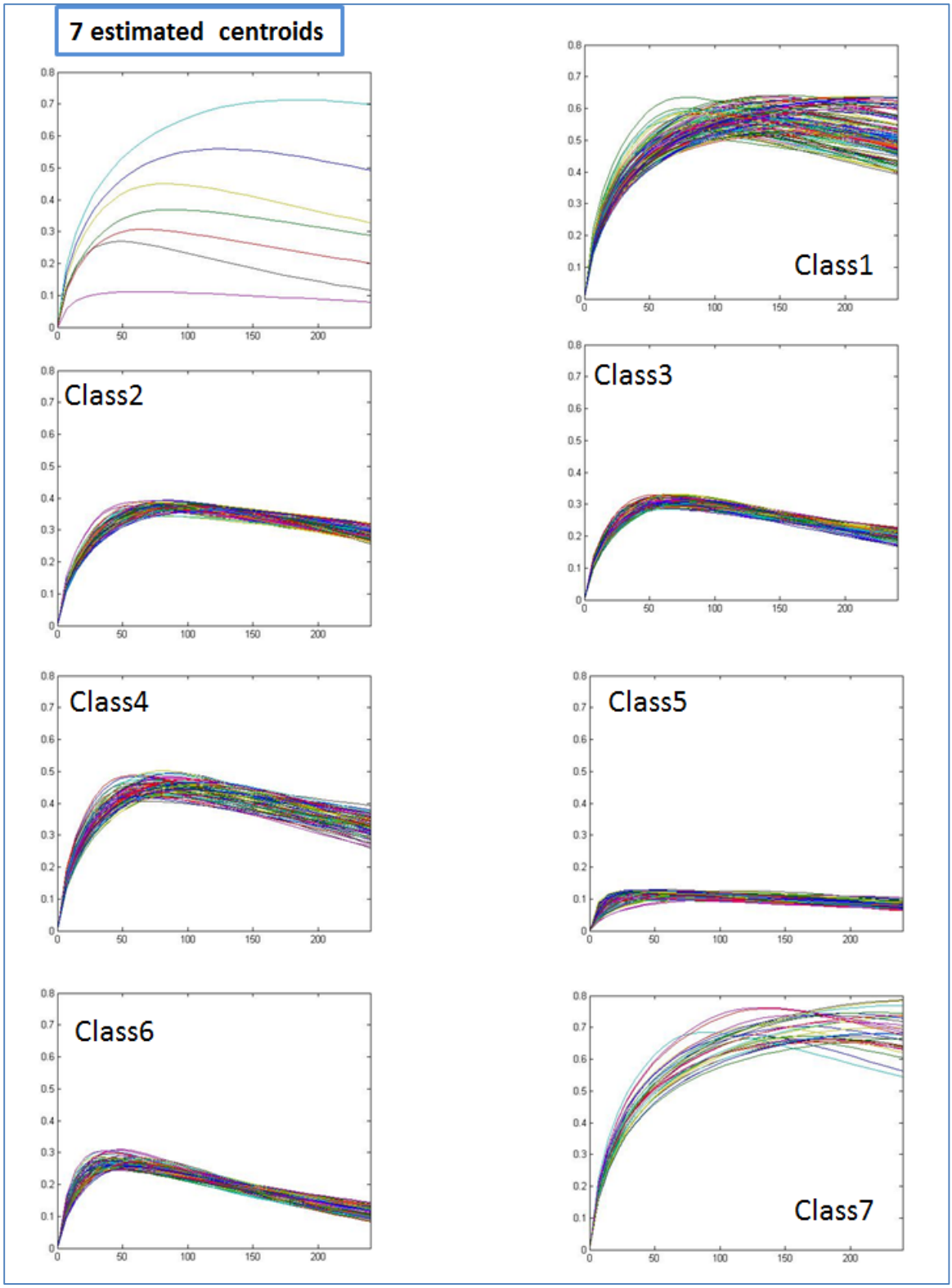


Figure 22. Qualitative representation of the estimated IBSL curves regarding the implementation of Case 3

### ✓ Case 4: Self-organization of data with self-evaluation of the number of classes

In this case, we examine the potential for fully automated data organization with self-evaluation, without any prior knowledge, neither on the number of principal classes nor on the nature of their centers. Firstly, we repeatedly apply the k-means algorithm to produce a population of class centers and automatically organize them in a set of (500, 29) centroids, performing k-means clustering with k=10 for 50 iterations calculating (10,29) centroids each time. Secondly, a Mean Shift scheme is performed for this population (500, 29) without any knowledge on the number of classes and the optimal centroids (8, 29) are automatically estimated based on the Silhouette criterion. In Figure 23, is illustrated the labeled centroids with respect to the estimated ones. Moreover, the above mentioned hypothesis/suspicion that additional, hidden information may exist within the given dataset and reference knowledge on the tissue characterization.

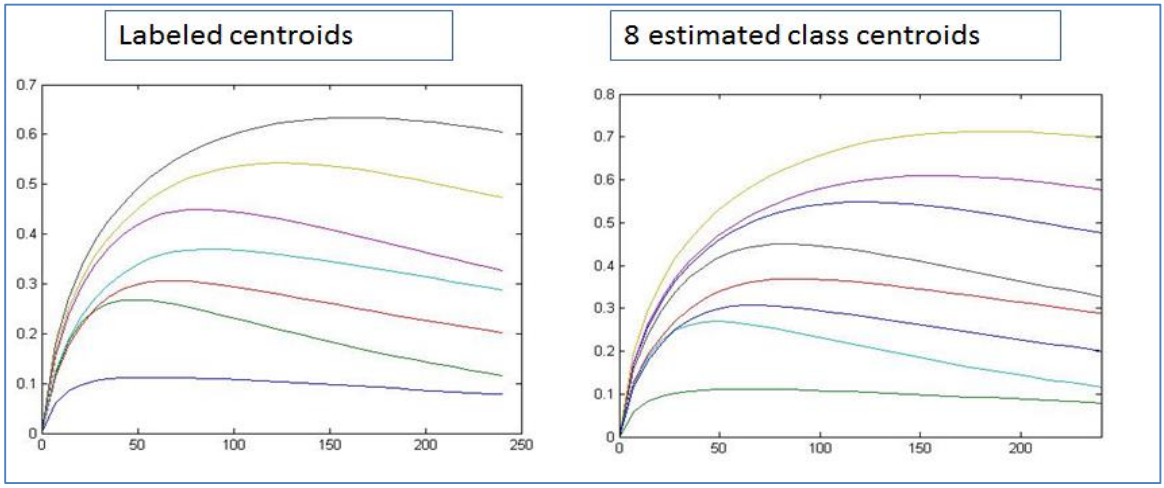


Figure 23. Labeled centroids (left) and estimated ones (right) representing the different tissue states.

The result of Case 4 is depicted in Figure 24, where the heavily mixed class 7 was efficiently splitted in two similar sub- classes. Furthermore, In Figure 25 is illustrated comparing the reference labels per class with the one derived via the proposed methodology utilizing the Euclidean metric as the distance function of k-means clustering. We may notice that reference class 7 appears to have been efficiently splitted into two sub-classes, probably revealing an extra clinical status of the tissue respect the originally defined ones.

Theses indications are very promising that the proposed algorithm sustains the ability to discriminate smaller population patterns within a possibly correlated environment. In essence, the proposed self-organization approach can indicate the existence of meaningful subclasses in the data.

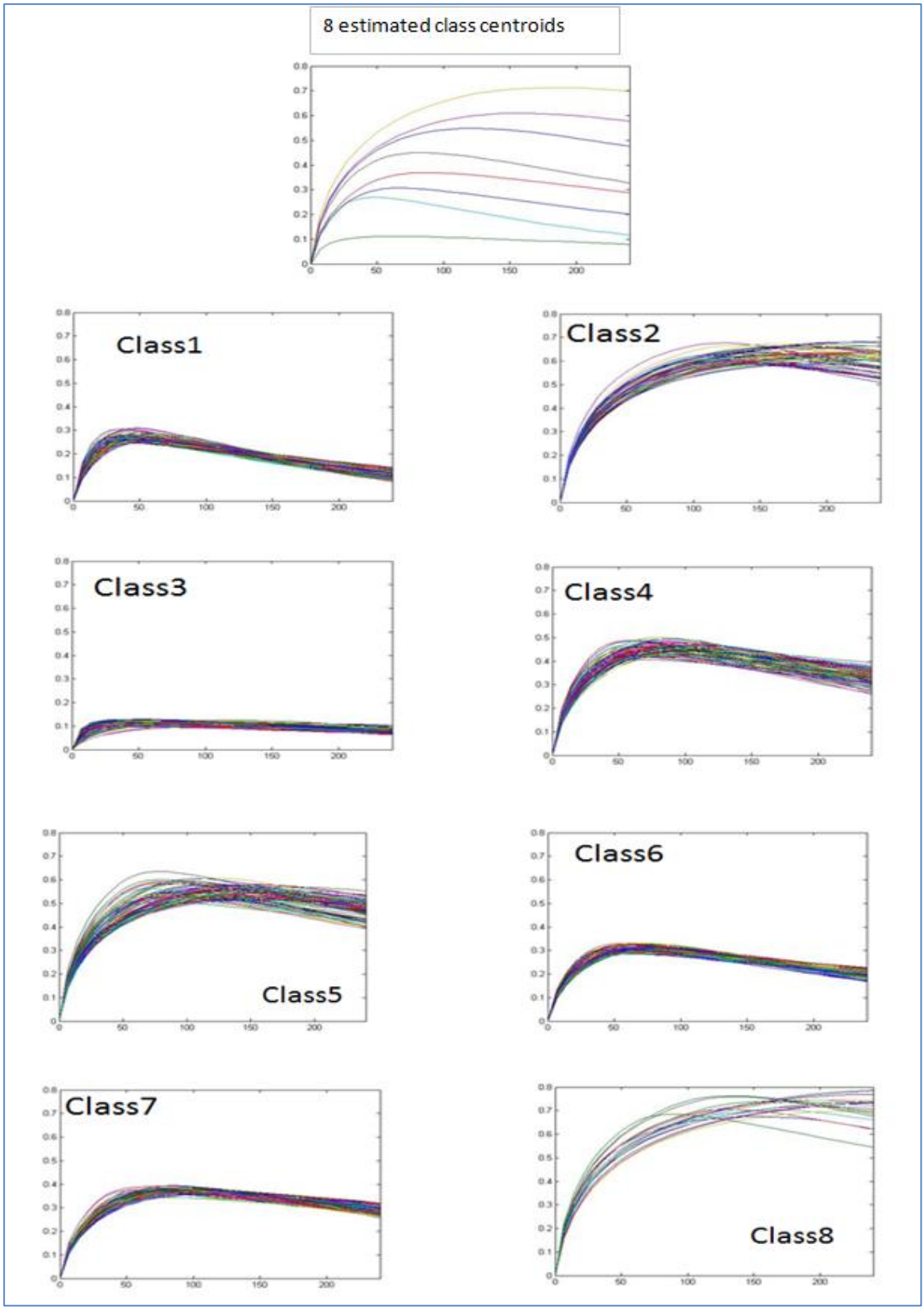


Figure 24. Qualitative representation of the 8 estimated IBSL classes regarding the implementation of Case 4.

***Classification Results (derived labels per class)***

*Reference labels per class*

| CLASS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | count |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 0 | 67 | 4 | 0 | 71 |
| 7 | 0 | 0 | 0 | 0 | 0 | 7 | 46 | 18 | 71 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| count | 71 | 71 | 71 | 71 | 71 | 74 | 50 | 18 | |

Figure 25. Confusion matrix reference and estimated labels utilizing the Euclidean distance metric within the k-means clustering procedure.

The estimated 8 cluster centroids via the MSH approach are applied to the testing, big dataset of 100000 samples, by performing Squared Euclidean Distance, constructing the new confusion matrix based on the ground truth labels given. It is important to mention that this dataset along with the reference classification originate from simulation procedures based on models validated on clinical data. The quantitative evaluation of applying Case 4 to this new dataset is depicted in Figure 26, via the construction of confusion matrix.

***Classification Results (derived labels per class)***

*Reference labels per class*

| class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | count |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 33808 | 290 | 0 | 0 | 0 | 0 | 0 | 0 | 34.098 |
| 2 | 56 | 18815 | 20 | 0 | 0 | 0 | 0 | 0 | 18.891 |
| 3 | 0 | 1443 | 20144 | 0 | 0 | 0 | 0 | 194 | 21.781 |
| 4 | 0 | 0 | 402 | 1643 | 0 | 0 | 0 | 15028 | 17.073 |
| 5 | 0 | 0 | 0 | 3678 | 58 | 0 | 0 | 0 | 3.736 |
| 6 | 0 | 0 | 0 | 1 | 3161 | 640 | 0 | 0 | 3.802 |
| 7 | 0 | 0 | 0 | 0 | 0 | 349 | 270 | 0 | 619 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ------- |
| count | 3364 | 20548 | 29566 | 5322 | 3219 | 989 | 270 | 15222 | |

Figure 26. Quantitative illustration via confusion matrix, applying Case 4 to dataset of 100000 simulated classes.

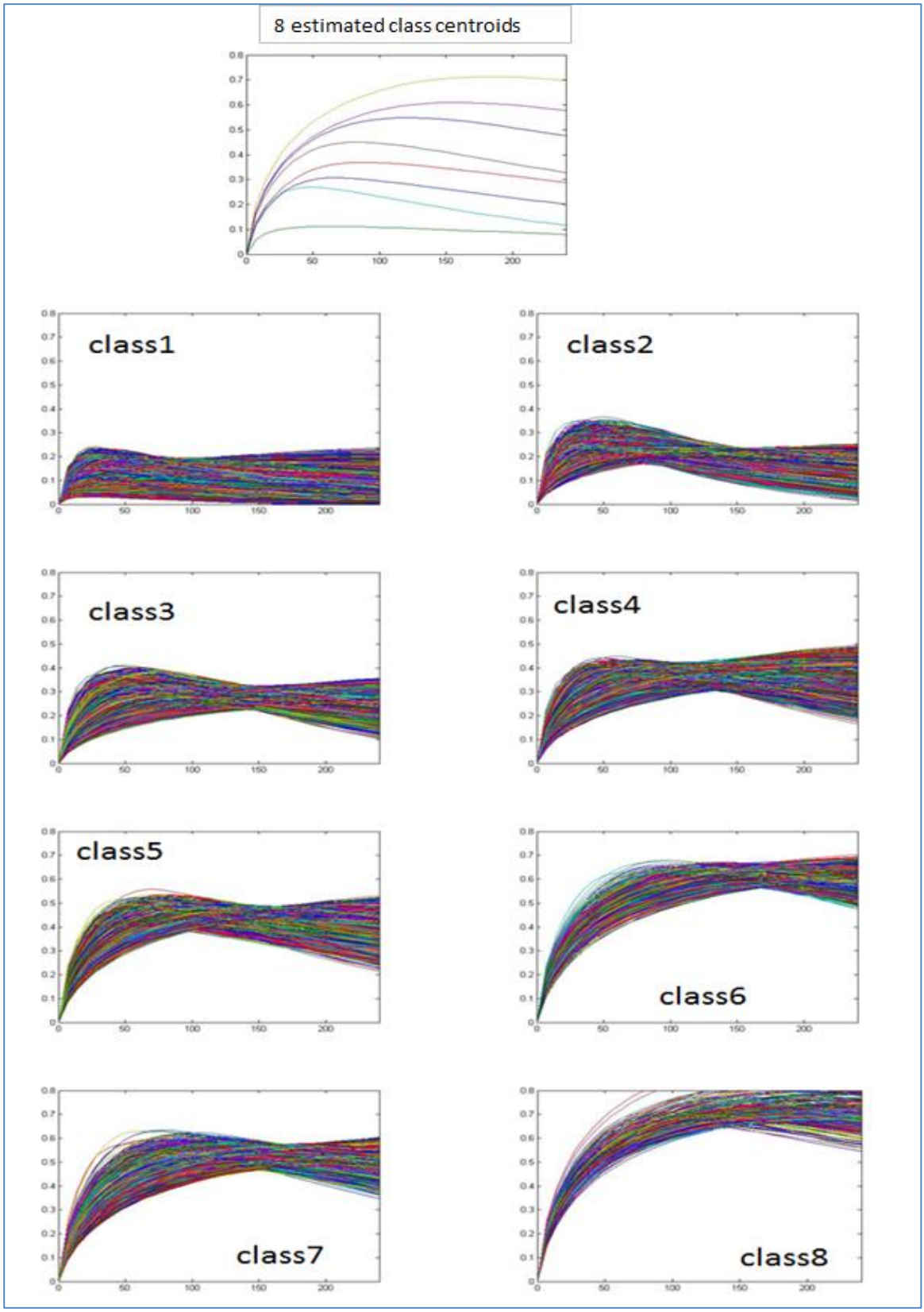


Figure 27. Qualitative illustration of 8 estimated classes applying Case 4 to dataset of 100000 simulated classes.

While the results of applying Case 4 to the training set (497 samples) presents a better distribution in confusion matrix and more efficiently separated curves based on shape similarities, in the case of the testing dataset (100000 samples) as it is depicted in Figure 27, the problem of accurate grouping still remains, producing heavily mixed classed. Thus, the new 8 centroids seem to improve the attitude of 497 samples but do not prove efficient for the overall classification metric sustains of curves but misses information based on pattern similarity.

*Testing alternative distance metric: Squared Euclidean Cosine*

Based on this observation, we examined the contribution of the cost function to the classification efficiency by adopting the Squared Euclidean Cosine distance metric, as introduced in (1.14) and performing the procedure so as to compare the two approaches and decide on the most accurate grouping of curves. The results after the application of this approach to the training and testing dataset are illustrated in Figure 28 and Figure 29 respectively, demonstrating that the newly adopted distance metric leads to bad separation of data based on magnitude differences but proves capable of identifying similar patterns. Furthermore, this result applying on the testing dataset of 100000 samples, as it is depicted in Figure 30, formed classes with heavily mixed problem in magnitude.

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Count |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 18 | 2 | 7 | 2 | 0 | 5 | 0 | 37 | 71 |
| 2 | 0 | 48 | 0 | 0 | 23 | 0 | 0 | 0 | 71 |
| 3 | 56 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 71 |
| 4 | 3 | 0 | 33 | 0 | 0 | 35 | 0 | 0 | 71 |
| 5 | 15 | 0 | 17 | 0 | 0 | 39 | 0 | 0 | 71 |
| 6 | 0 | 0 | 21 | 42 | 0 | 5 | 3 | 0 | 71 |
| 7 | 0 | 0 | 5 | 26 | 0 | 2 | 38 | 0 | 71 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - |
| count | 92 | 51 | 53 | 70 | 23 | 97 | 41 | 36 | |

Figure 28. Confusion matrix estimated classes applying Case 4 to the data set of 497 clinical samples utilizing the Squared Euclidean Cosine distance metric.

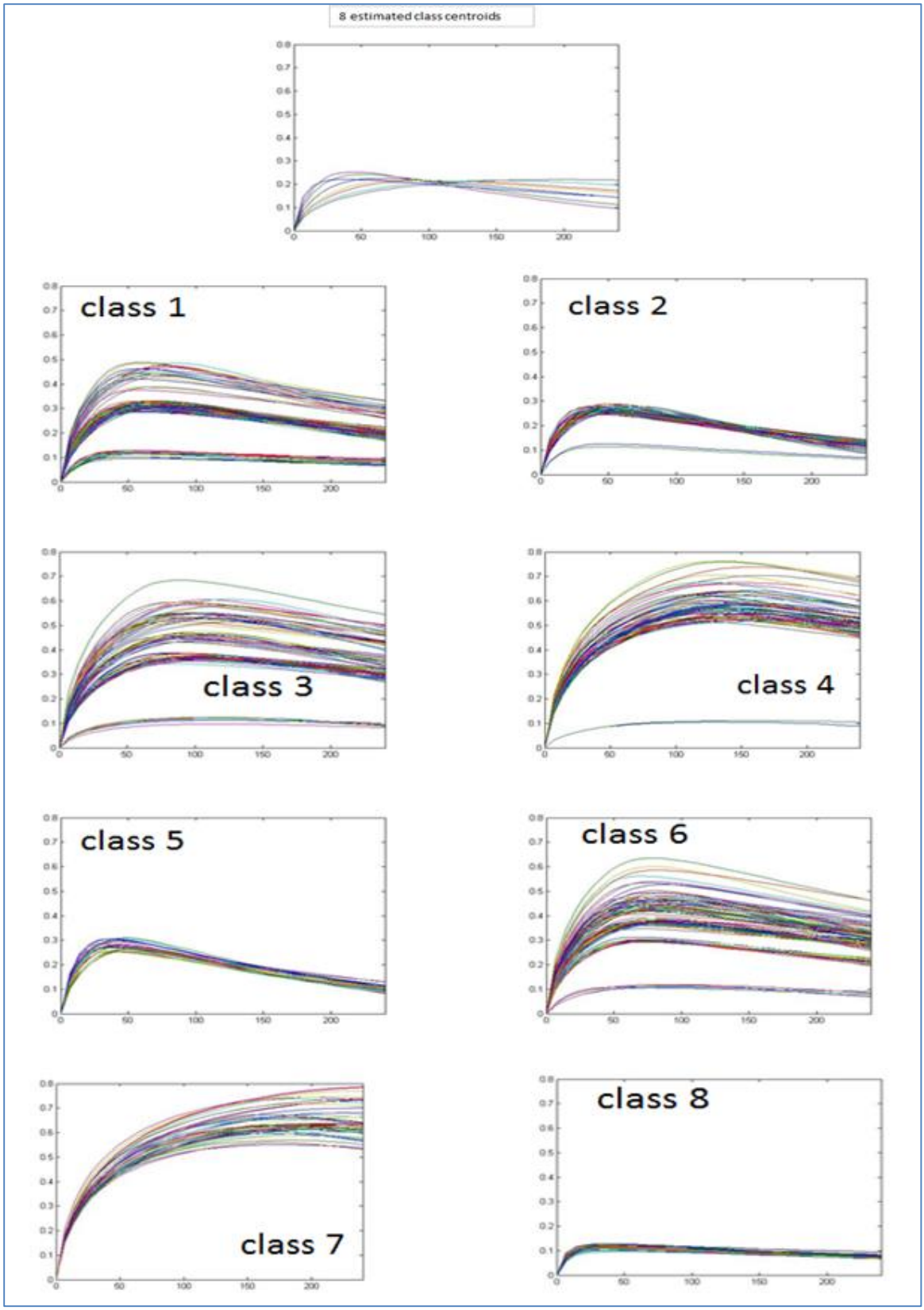


Figure 29. Qualitative illustration of the 8 estimated classes applying Case 4 to the dataset of 497 clinical samples utilizing the Squared Cosine distance metric.

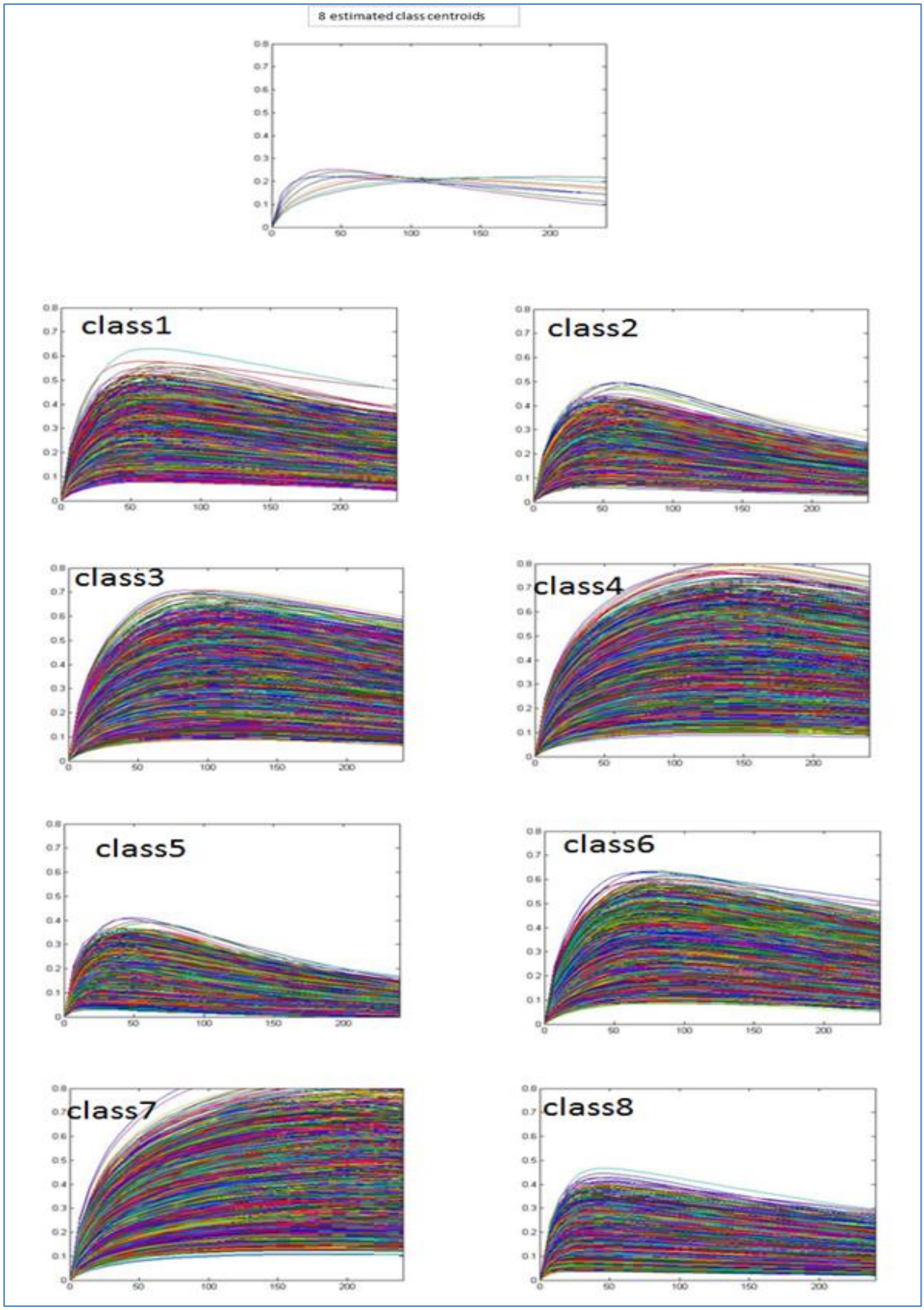


Figure 30. Estimated classes applying Case 4 to the dataset of 100000 simulated samples utilizing the Squared Euclidean Cosine distance metric.

### ➢ Conclusions of Self-organization of data with self-evaluation of the number of classes approach

In scenario 1, we introduce a novel approach for automatically organizing data by fusing clustering and resampling approaches. Our methodology was applied to biomedical data aiming at efficiently classifying characteristic intensity curves of Backscattered Light (IBSL) representing cervical tissue to one of seven reference labels of clinical importance. Both qualitative and quantitative preliminary results validate the efficiency of the proposed technique to search for hidden information and detect statistical knowledge from unknown environments, giving rise to its utilization as an efficient tissue-characterization tool. The proposed approach indicates an extra clinical status of tissue in addition the originally defined ones and show good potential in discriminating smaller population of patterns within a possibly correlated environment.

Based on this approach two significant indications reveal. Firstly, the results demonstrate, the inability to efficiently self-organize and classify the testing dataset into well discriminated and meaningful groups of curves. Towards this direction, a more intuitive and advance approach should be considered. This technique needs generalization in order to apply to completely unknown environments and to be efficient in big data set. Secondly, after the examination of two distance metrics it was arisen, that the Euclidean distance classifies by the magnitude but loose the shape, while the Squared Cosine distance gives priority to the shape. Consequently, the necessity of a new distance metric, which will combine the above important traits, is essential.

## 5.2 Scenario 2: Recursive k-means mode

The evaluation of the proposed methodology on the clinical dataset (497 IBSL versus time curves validated through actual biopsy) is based in the construction of the confusion matrix of estimated labels against the ground-truth ones, assigned through tissue biopsy and characterization. The confusion matrix encodes the distribution of reference curves from each class label over the estimated classes. The confusion matrix is a representative means, in the form of a table, to examine the distribution of reference curves from each class label over the estimated classes. Each row of the table reflects the corresponding label and each column represents the estimated class associated with the corresponding class center. Thus each array element $(i, j)$ indicates the number of curves from labeled (ground-truth) class $i$ that have been assigned to estimated class $j$ using a fixed set of class centers.

### ✓ Case 1: Iterative k-means initialization with clinical centroids

Utilizing the 7 clinically approved ground truth curves as initial seeds for the proposed clustering recursive-mode k-means clustering of the clinical set of 497 samples, produces the results illustrated in Figure 31. The confusion matrix clearly reveals a remarkable efficiency in properly discriminating the estimated classes, apart from the last one, which seems to have dispersed in classes 6 and 7. This is a confirmed by the qualitative results in Figure 32, where estimated classes 1-5 seem to sustain a very compact form, while classes 6-7 pertain sample of slightly varying shape and amplitude.

***CLASSIFICATION RESULTS ( derived labels per class)***

| *Reference labels per class* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CLASS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 71 |
| 7 | 0 | 0 | 0 | 0 | 0 | 25 | 46 | 71 |
| count | 71 | 71 | 71 | 71 | 71 | 96 | 46 | |

Figure 31. Confusion matrix illustration, regarding the implementation of recursive mode k-means, initialized by the expert knowledge and applied to the clinical set for 497 IBSL.

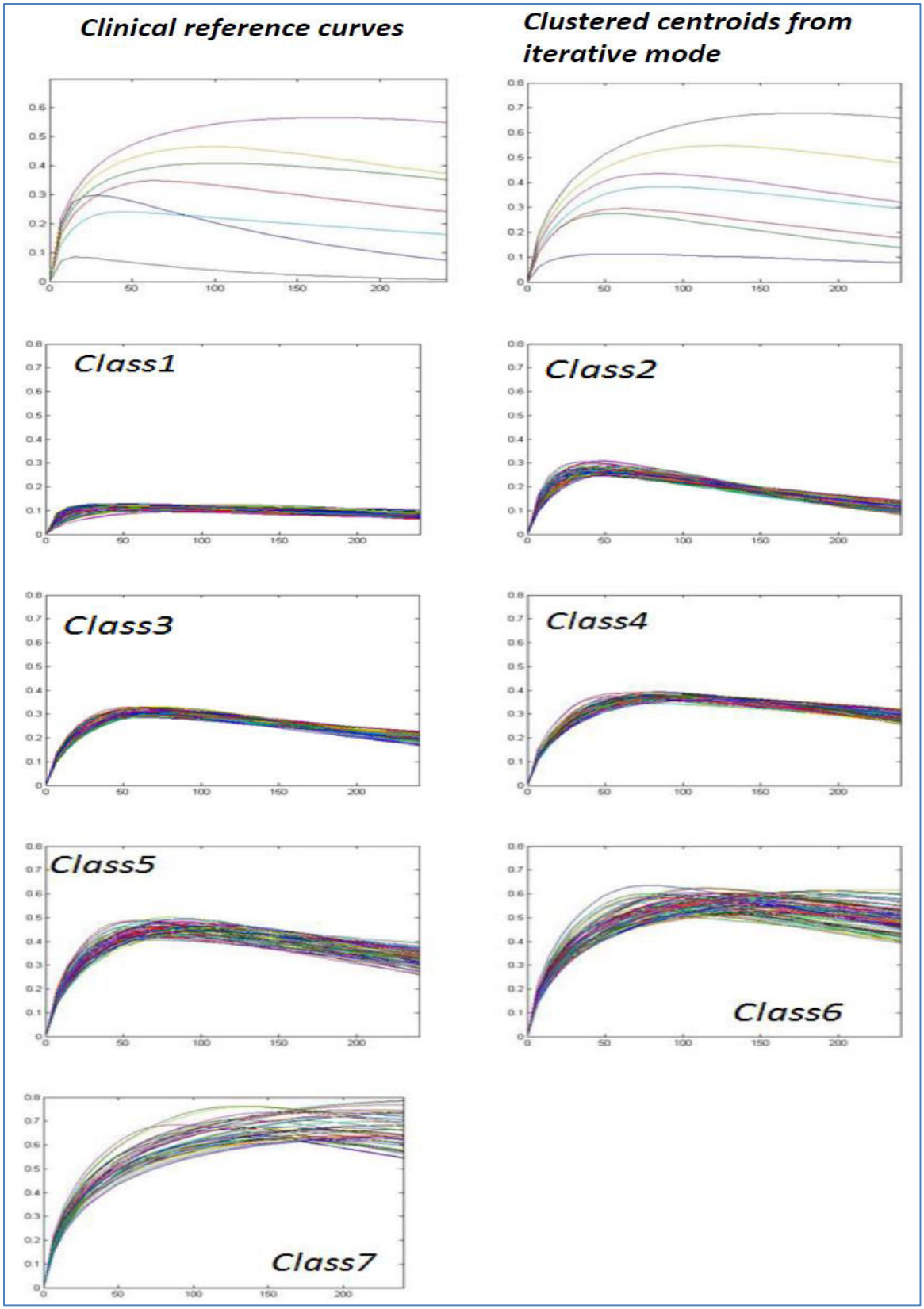


Figure 32. Calculated clusters which are group based on recursive k-means mode, initialized by the expert knowledge are applied to the clinical set 0f 497 labeled IBSL curves.

### ✓ Case 2: Iterative k-means initialization with random seeds

In this section, we examine the results of clustering scheme initialized by random seeds. In this scheme, the initial centroids will be collected randomly by k-means without external interventions. The following procedure remains the same by following the steps as we described above. Random initialization seems to have a deteriorating effect on the compactness of the estimated clusters and the consistency between the reference and the estimated centroids. Thus, the advantage of exploiting expert knowledge as in the proposed scheme is validated both quantitatively and qualitatively.

The corresponding results of the proposed method are depicted in Figure 33 where a confusion matrix is revealing and the Figure 34, where the estimated classes are illustrated. The confusion matrix obviously depicts a good efficiency in distinguishing the estimated classes, although it seems to have scattered on classes 6, 7 mainly and 5 slightly. It is clear from confusion matrix that some of the classes are mixed. Also, as can been seen from the Figure 33 the estimated classes seem to lose their compactness and they are not clearly distinguished one from another. Furthermore, the estimated classes are depicted mixed. Specifically, the class 4 seems to contain two different types of curves, losing the magnitude. This is a confirmed by the qualitative results in Figure 34 where estimated classes pertain sample of slightly different curvature for the reference classes 1, 2, 6.

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 70 | 1 | 0 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 64 | 7 | 0 | 71 |
| 7 | 0 | 0 | 0 | 0 | 6 | 40 | 25 | 71 |
| count | 71 | 71 | 71 | 141 | 71 | 47 | 25 | |

Figure 33. Confusion matrix depiction based on recursive mode k-means, initialized by random seeds, reveals the quantitative estimation of resulted clusters.

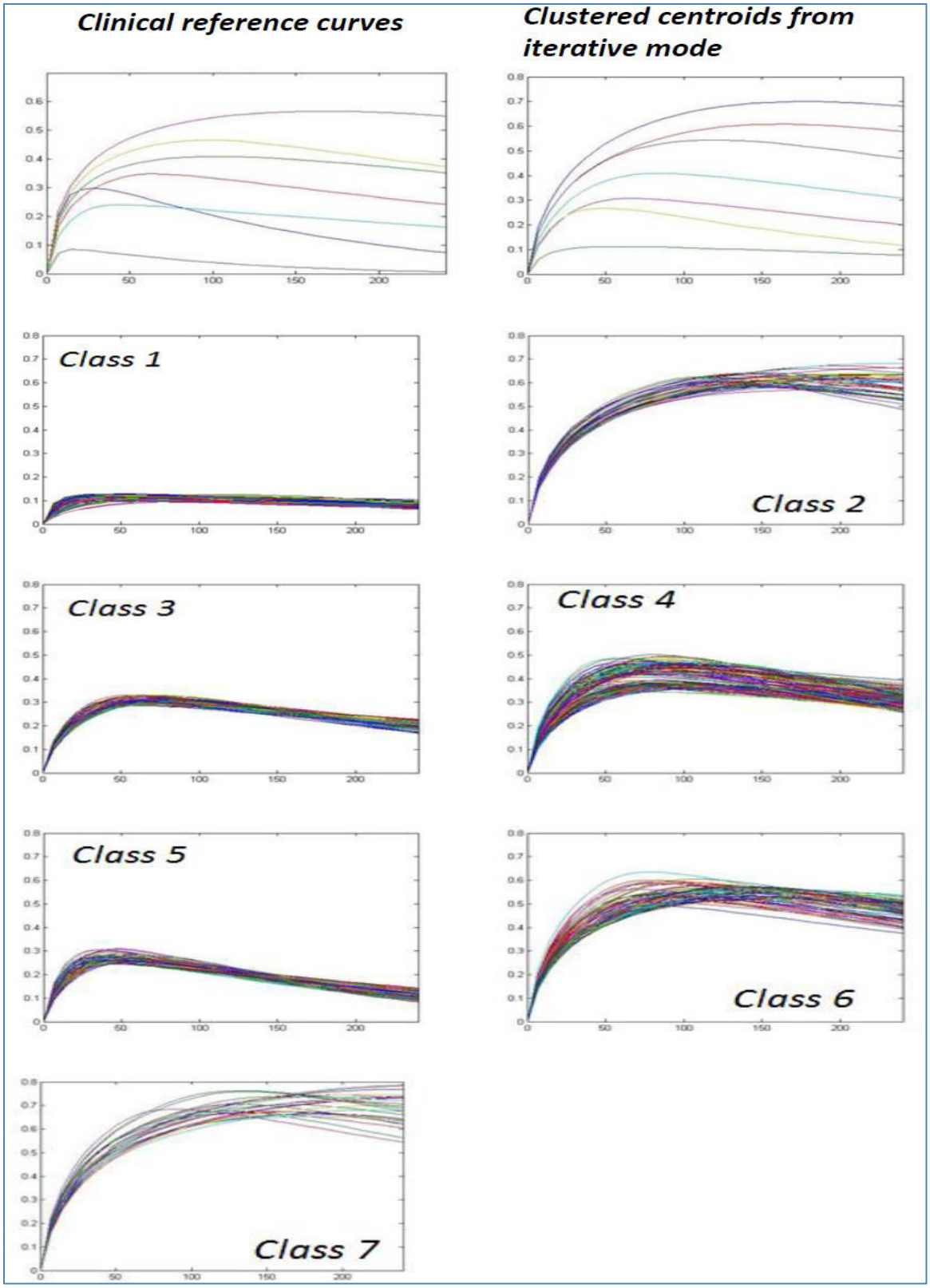


Figure 34. Qualitatively illustration of estimated classes, regarding the implementation of recursive mode k-means with random seeds initialization and applied to the clinical set of 497 labeled IBSL curves.

### ✓ Case 3: Conventional k-means mode with clinical centroids initialization

With the view to compare and rank our recursive mode k-means clustering methodology, where the extracted centroids are innovately improved with respect to their current and previous value, we also consider the classic k-means approach on the clinical dataset with similar initialization strategies. Two aspects are followed, similar to above methodology in order to produce clusters comparing to the previous estimated clusters. The first implementation of k-means is with expert knowledge initialization and the second implementation of k-means is with random initialization.

The corresponding results are represented in Figure 35 and Figure 36, indicating the outperformance of the proposed weighted center improvement scheme against the fixed update of the classic k-means implementation. The confusion matrix reveals non-compact classes. Also as can be seen from both Figure 35 and Figure 36, the extracted classes contains curve of different shape and magnitude. The resulted clustered centroids which are illustrated in Figure 36 are not well separated. Without the refinement of the extracted centers, sample data are dispersed into mixed and classes. It is also noteworthy that the utilization of initial points imposed by the expert knowledge enhances the accuracy and robustness of the final cluster centers in the classic k-means approach.

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 32 | 0 | 39 | 0 | 0 | 71 |
| 6 | 0 | 0 | 2 | 0 | 0 | 69 | 0 | 71 |
| 7 | 0 | 0 | 0 | 0 | 0 | 24 | 47 | 71 |
| count | 71 | 142 | 34 | 71 | 39 | 93 | 47 | |

Figure 35. Confusion matrix representation based on implementation of classical k-means mode, initialized by the expert knowledge and revealing the classification of the estimated clusters.

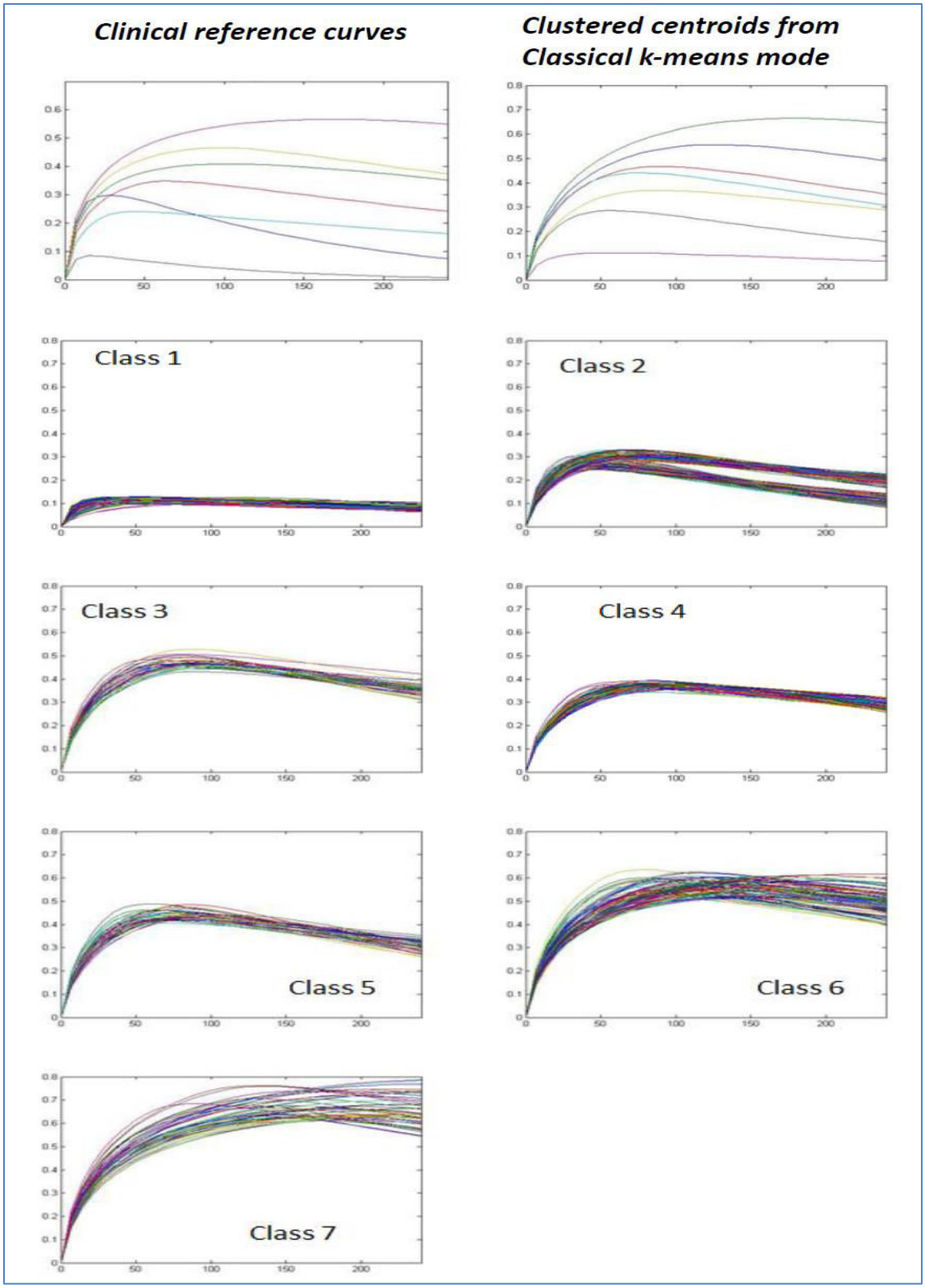


Figure 36. Qualitative representation of estimated clusters regarding the implementation of classical k-means, initialized by the expert knowledge and applied to the clinical set of 497 labeled IBSL curves.

### ✓ Case 4: Conventional k-means with random seeds initialization

In order to continue with the comparison of results, we proceed to the second implementation of classical k-means approach, with random initialization. The produced results, which are depicted in Figure 37 and Figure 38 revealing the heavy mixing problem in classes. The formed clusters are represented composite. Moreover, different types of classes are blended without keeping their unique form. In addition, the resulted clusters of this approach are not separated clearly as can be seen from the Figure 37.

In this method the mixing problem is heavier than in previous aspect with random initialization, where the final clusters are more accurate and robust. Consequently, we can notice that the expert knowledge enhances the accuracy and robustness of the final clustered centers, in Recursive k-means scenario and Classical k-means scenario. From the above graphs is clear that the expert knowledge has a significant contribution on implementation of k-means. Also, the proposed method indicates the outperformance of our proposed methodology against the replacement methodology of Classical k-means implementation

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | count |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 36 | 35 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 71 |
| 7 | 0 | 0 | 0 | 0 | 0 | 24 | 47 | 71 |
| count | 71 | 142 | 36 | 35 | 71 | 96 | 47 | |

Figure 37. Quantitative distribution of estimated classes based on performance of classical k-means, initialized by the random initialization.

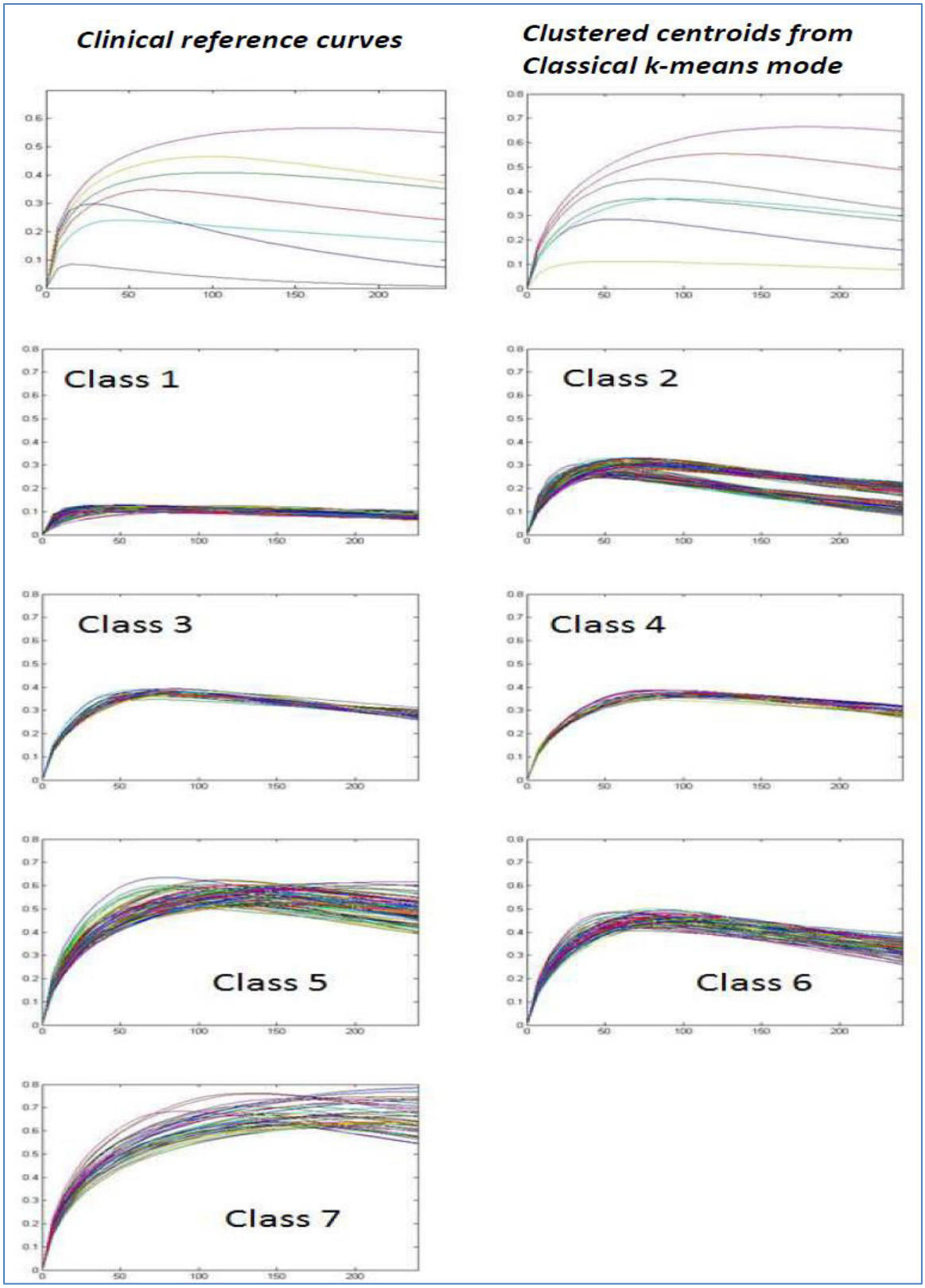


Figure 38. Qualitative illustration the calculated classes regarding the implementation of classical k-means initialized by random seeds and applied to the clinical set of 497 labeled IBSL curves.

## ✓ Conclusions of Recursive k-means approach

In this scenario we propose a novel algorithmic scheme for self-organizing data, adopting a Recursive k-means mode in order to extract adequate information of the data and classify cervical cancer AW sample into meaningful classes. The stability of k-means clustering through multiple initializations was examined in this scenario. Also, a novel approach for automatically organizing data by fusing information of data distribution and expert knowledge was introduced. Moreover this method is efficient to search for hidden information and detect statistical knowledge from completely unknown environments, as it is shown in Recursive k-means initialized by random seeds. The examined methodology indicates that the utilization of initial points by the expert knowledge enhances the accuracy and robustness of the final centers. We utilize both expert and experimental evidence, by means of appropriate clustering criteria resulting in a recursive relaxation of class centers.

Our approach was applied to biomedical data aiming at efficiently classifying characteristic intensity response curves from cervical tissue to seven reference labels of clinical importance. As can be seen from the Figure 39 the proposal recursive k-means approach based on initialization with expert knowledge, provides an efficient classification of the experimental data, revealing a remarkable efficiency in properly discriminating the estimated classes.

The resulted clustered centroids are illustrated clearly distinguished one from the other, creating compact and well separated clusters based on shape and magnitude. In comparison with both approaches of conventional k-means, the proposal k-means approaches (expert knowledge, random seeds), dominate on first ones via production refinement centroids closed to reference classes. Both qualitative and quantitative preliminary results validate the efficiency of recursive k-means case initialized by reference seeds to search for hidden information and detect statistical knowledge from unknown environments starting from biological grounds, giving rise to its utilization as an efficient tissues characterization tool.

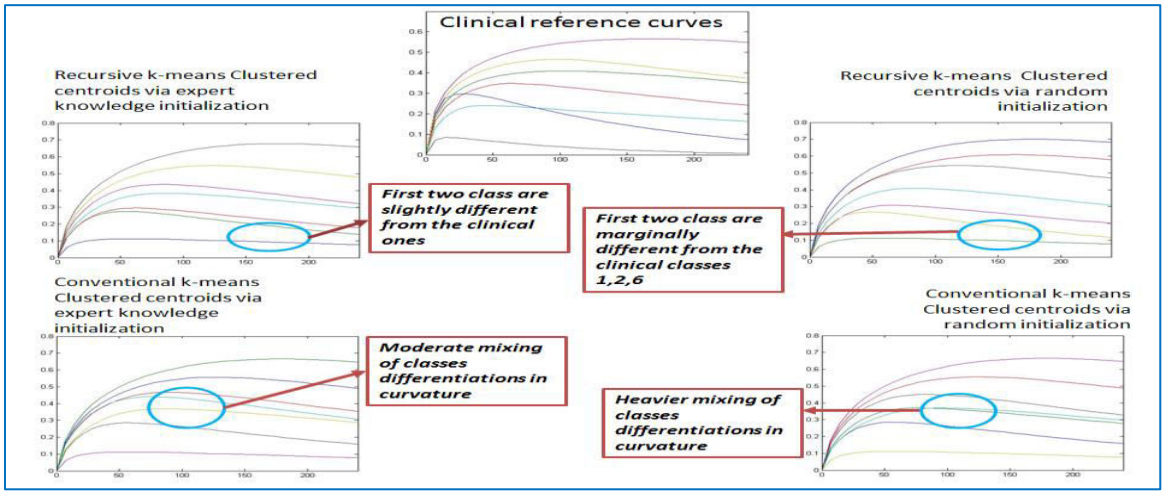


Figure 39. Qualitative presentation of comparative results based on recursive k-means implementation and the conventional k-means mode, versus of clinical curves.

## 5.3 Scenario 3:Exploration of Distance metric and Class formation Strategy Testing Results

### 1. *Efficient Assessment of sequential clustering*

Following and expanding the previous scenario, we examine the effectiveness of a new distance strategy through the sequential operation of Squared Euclidean and Squared Euclidean Cosine metric in order to capture both size and shape characteristics of the dominant classes and test if the resulting cluster centers can yield efficient unmixing of the large dataset. We test different We examine different distances for organization of both the labeled (497 curves) and the larger dataset (100000 samples), adopting, first the 7 reference clinical curves as class centers and then examining the 7 mean of curves of the labeled classes in the training set (497 curves). The process follows two distinct steps using two distance metrics. We firstly apply the Squared Euclidean distance metric for the clustering of curves treated as unlabeled data and at a second stage we utilize k-means for each class with k=2 and the Cosine metric as distance function. Thus, we implement a splitting scheme resulting in 14 new but well separated clusters.

The summarization of centers from the proposed scheme is presented in Figure. 40, comparing the 7 ground-truth curves and the centroids of the 7 clinical states with the estimated 14 cluster centroids. It can be observed that the 14 curves may reveal several hidden sub-classes containing both pattern and magnitude differentiations. After clinical validation, some of these subclasses may reveal additional clinical states that could be utilized to characterize the grade of cervical neoplasia.

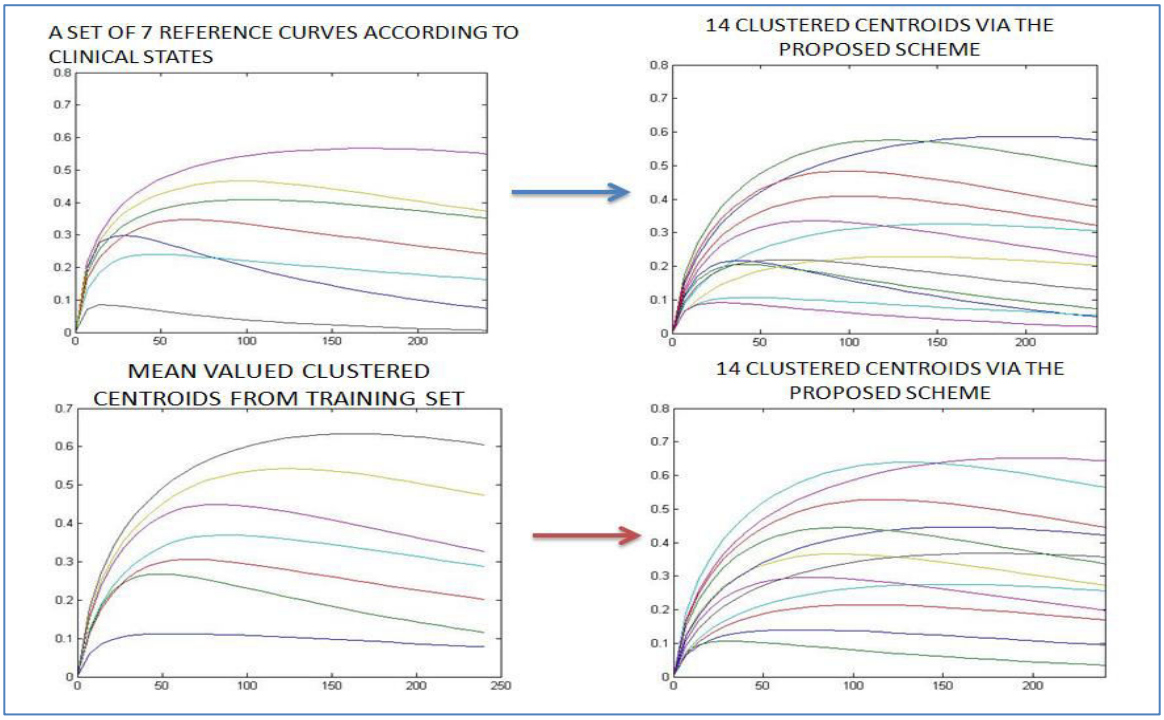


Figure 40. Reference and estimated tissue characterization curves (first row) along with labeled class means and estimated centroids (second row) utilizing a sequential distance metric.

In this scheme we consider as given the reference curves representing the clinical states of cervical tissue and the ground-truth centroids characterizing each class of the training (497 labeled samples). Following the evaluation of the sequential distance scheme, we attempt to cluster the large dataset of 100000 curves. We first cluster the testing set based on Squared Euclidean distance of each sample curve from the given clinical centroids. The results are summarized in Figure 41, demonstrating that neither set of clinical centroids is able to efficiently cluster the large testing set of unlabeled data.

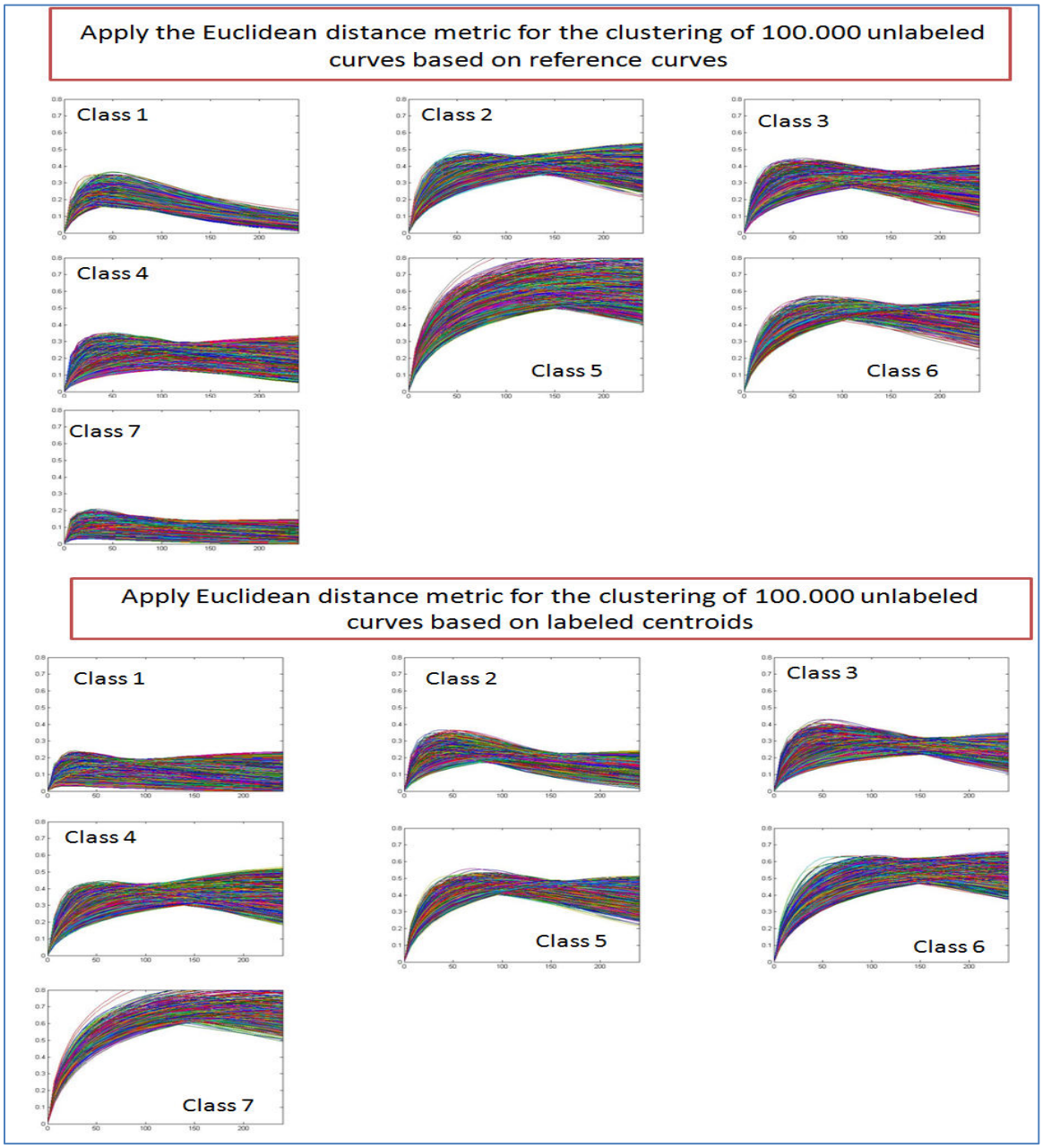


Figure 41. Qualitative results from clustering the testing set based on the Euclidean distance from the reference clinical curves (top section) and the labeled class centroids from the training set (lower section).

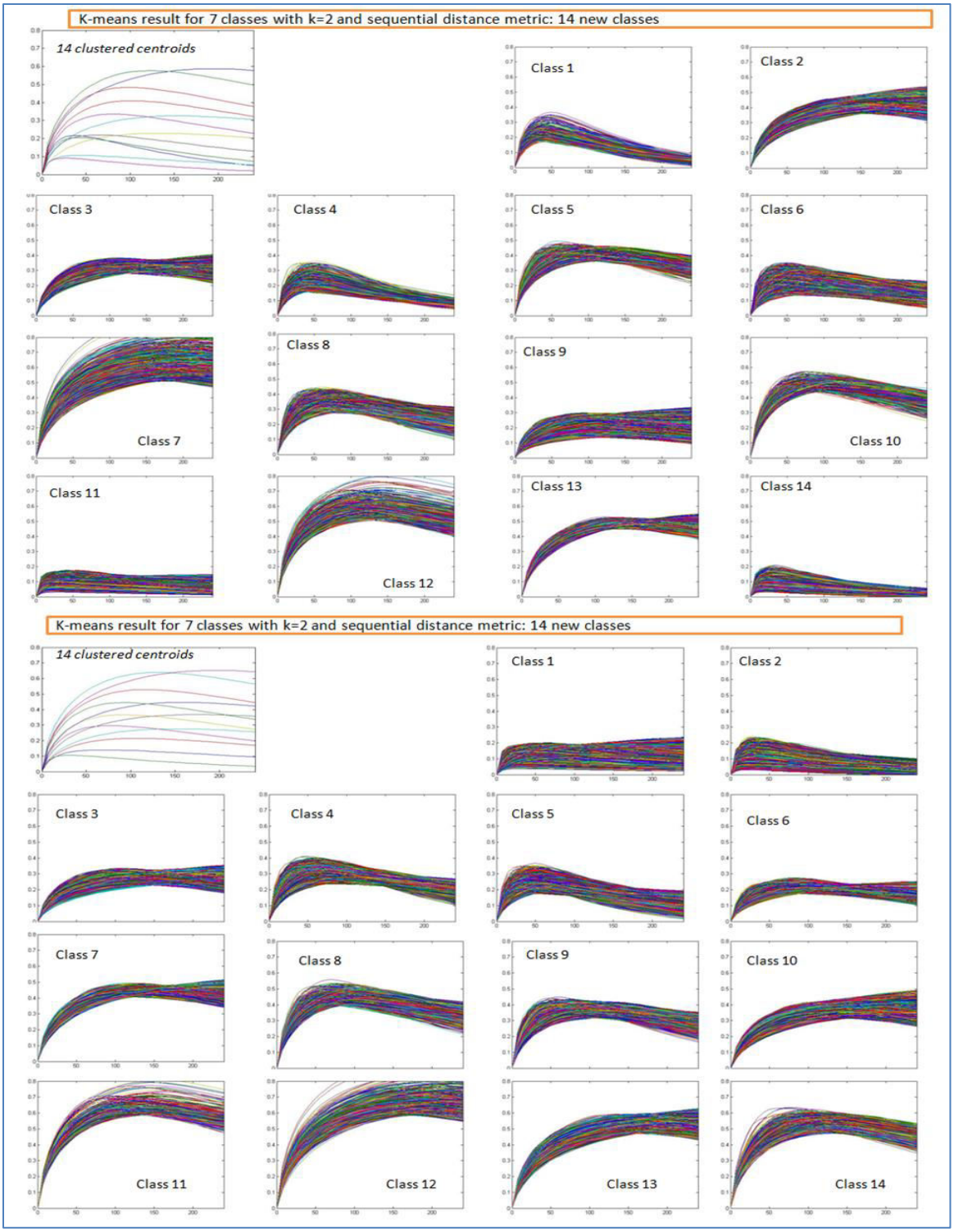


Figure 42. Clustering of testing dataset into 14 classes based on the sequential application of two distance metrics. The top section depicts the resulting 14 classes using the reference curves for Euclidean clustering and then k-means for Cosine clustering. The lower section illustrates the resulting 14 classes according the using of k-means based on 7 classes which are derived from the labeled centroids.

After the application of Euclidean distance metric to the testing data set, we use k-means with number of clusters k=2 and Squared Cosine distance in order to split each of the seven classes into new subclasses paying attention to the shape of curves. A qualitative representation of the estimated classes based on this sequential scheme is depicted above in Figure 42, yielding good separation by both shape and magnitude. The resulting 14 classes, in both cases, show good separation of the curve population. Consequently, we can infer good behavior of the sequential metric in the large data set as both Squared Euclidean Cosine and Squared Euclidean distances preserve their characteristics.

However, the center curves indicate a lot of overlapping, so that some of them could be merged. The process of merging and splitting are not clear. Thus, we proceed with another metric, which can be applied directly within the clustering process. This metric is described in the following section.

### 2. *Proposed Combined Distance Metric*

Building on the conclusions of the previous strategy, we evaluate the potential of the New combined distance metric forming the suitable combination of Squared Euclidean and Squared Euclidean Cosine metrics in order to perceive information based on both size and shape characteristics of the dominant classes. The rationale for examining yet another distance comes from the peculiarities of the dataset itself. Also we test if the basic clustered centers can yield efficient grouping unmixing of the dataset of 497 samples. As we mentioned before, we examine Squared Euclidean and Squared Euclidean Cosine for the organization of the labeled adopting, at first step the derived 8 clustered centroids which are produced according the implementation of scenario 1, respectively for each distance. In the second step, we examine the efficient self-organization of the 497 samples with self-evaluation of the number of classes, performing the previous scenario 1, applying in k-means the New Combined distance. Finally, we test how well the three above distances reflect the distribution of the 7 clinical reference centroids. We explore the possibilities of each of three distances to perform the testing set of IBSL curves in such scheme to perceive their traits in order to derive the corresponding class centers.

First, we apply the Squared Euclidean metric for the clustering of curves treated as unlabeled data. In Figure 43 is depicted the qualitative results regarding the implementation of scenario 1 and the 8 estimated centroids. In Figure 44 is illustrated the qualitative representation of the 497 IBSL curves based on their minimum distance of the 7 clinical reference centroids.

Secondly, we utilize the Squared Euclidean Cosine metric in order to organize the 497 curves based on the calculation of the minimum distance from the 8 produced cluster centroids by the implementation of scenario 1 based on this distance, as it is presented in Figure 45. Also in Figure 46, we test the efficiency classification of Squared Cosine distance metric for the 497 samples and the 7 clinical reference centroids.

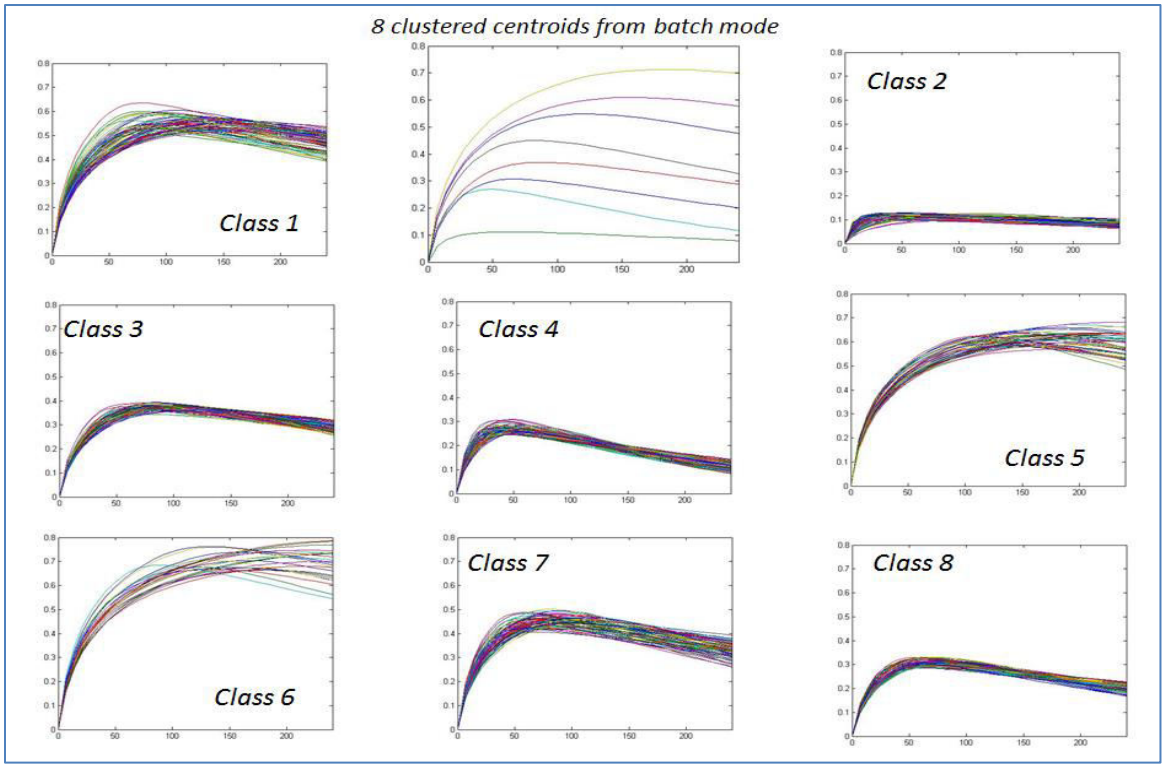


Figure 43. Qualitative representation of the 8 estimated IBSL curves by applying Euclidean distance regarding the implementation of scenario 1 and 8 produced clustered centroids.

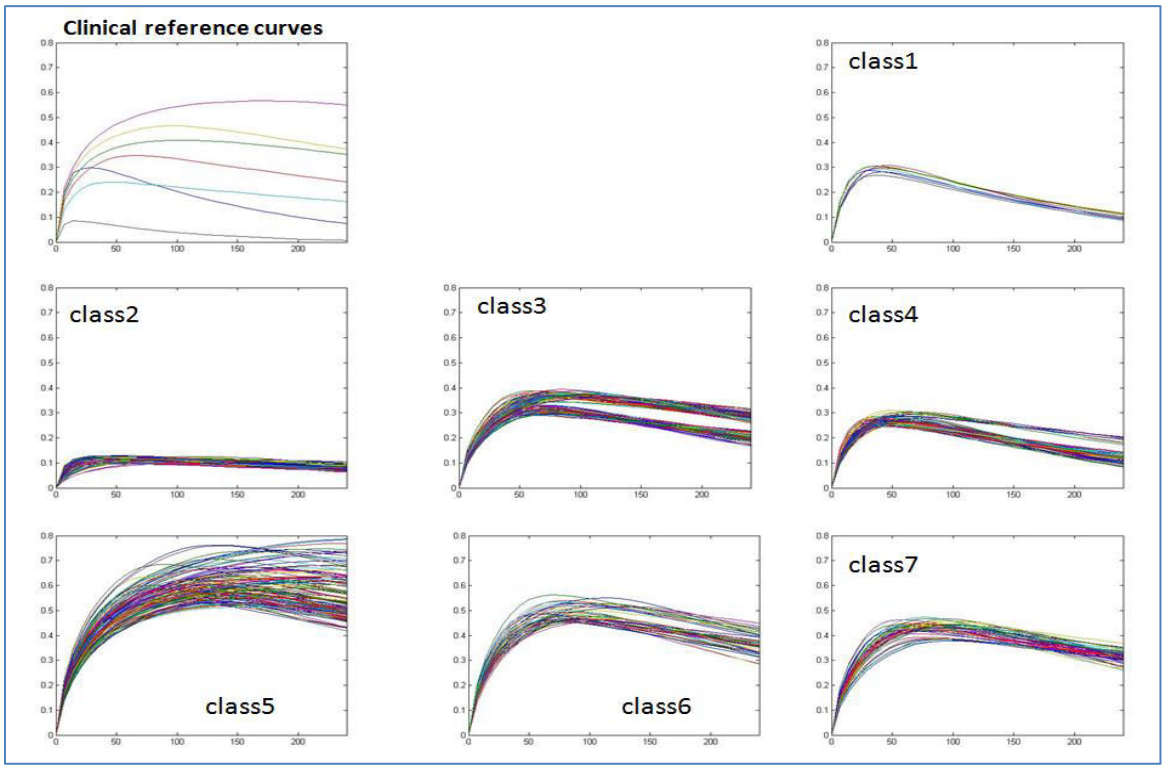


Figure 44. Data grouping of the experimental dataset based on the minimum Euclidean distance from the clinical reference curves.

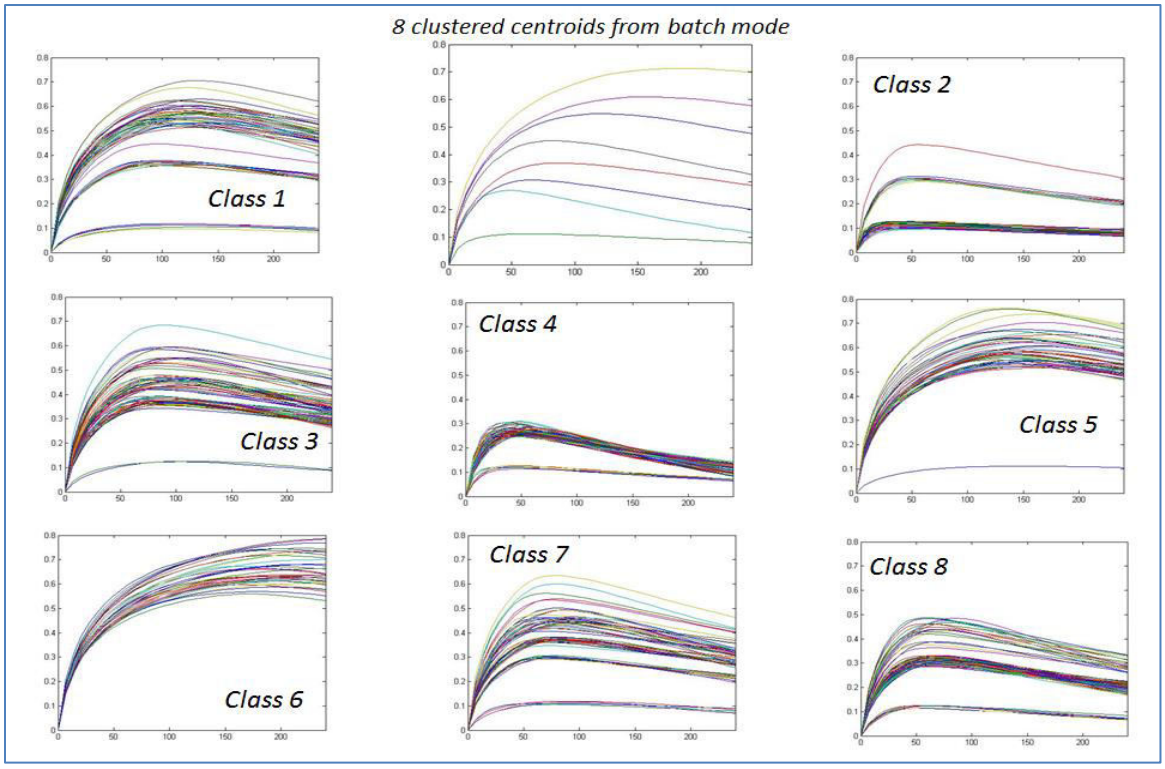


Figure 46. Estimated classes regarding the implementation of scenario 1 and Squared Cosine distance.

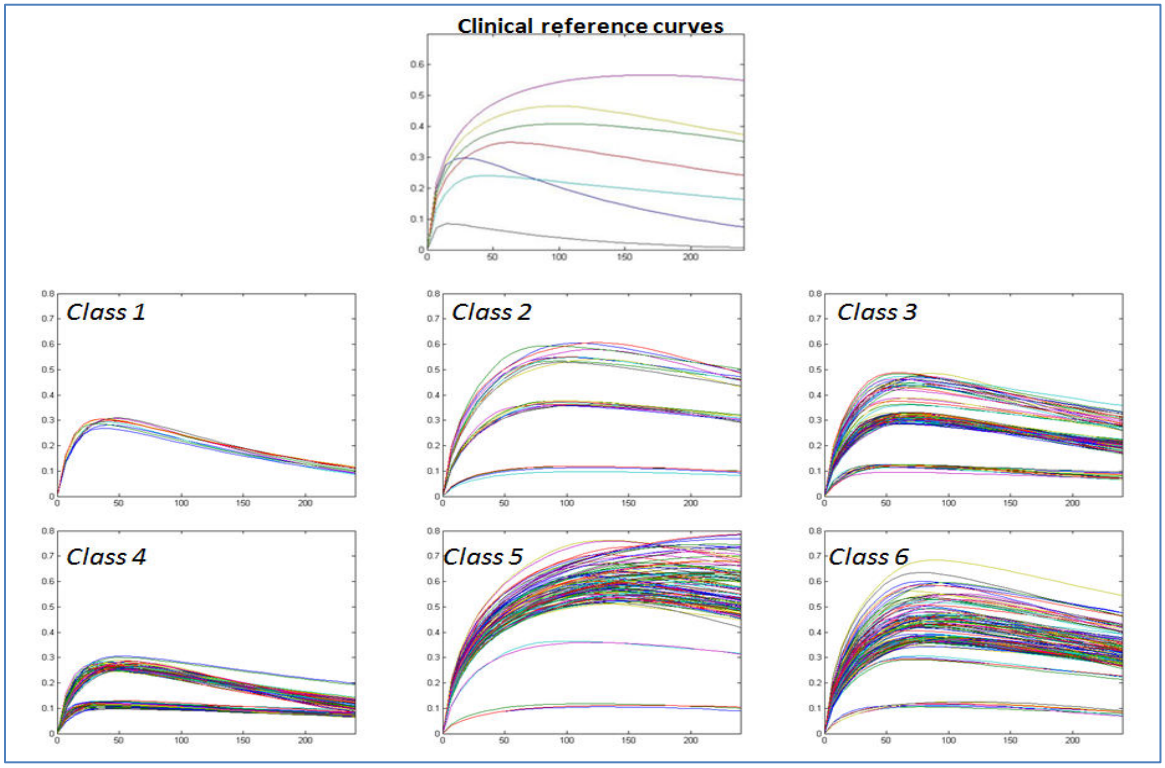


Figure 47. Data grouping of the experimental dataset based on the minimum Cosine distance from the clinical reference curves.

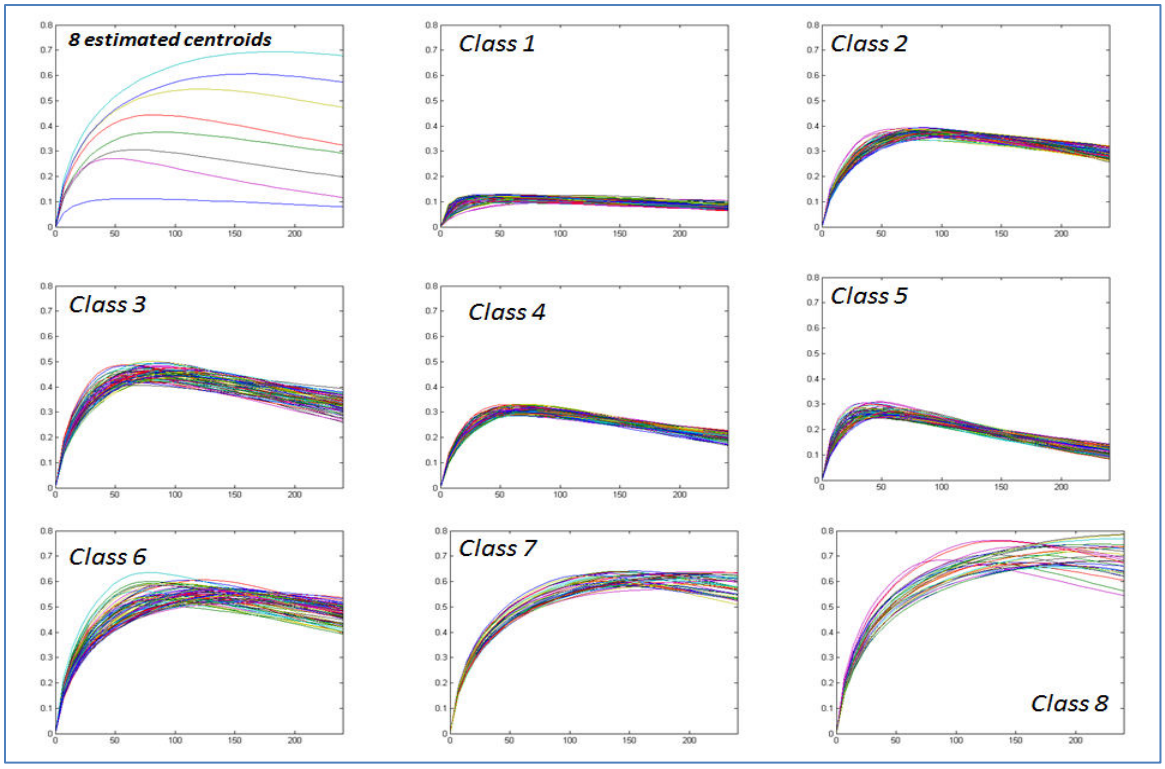


Figure 48. Qualitative representation of the 8 estimated IBSL curves by applying New combined distance regarding the implementation of scenario 1 and 8 produced clustered centroids.

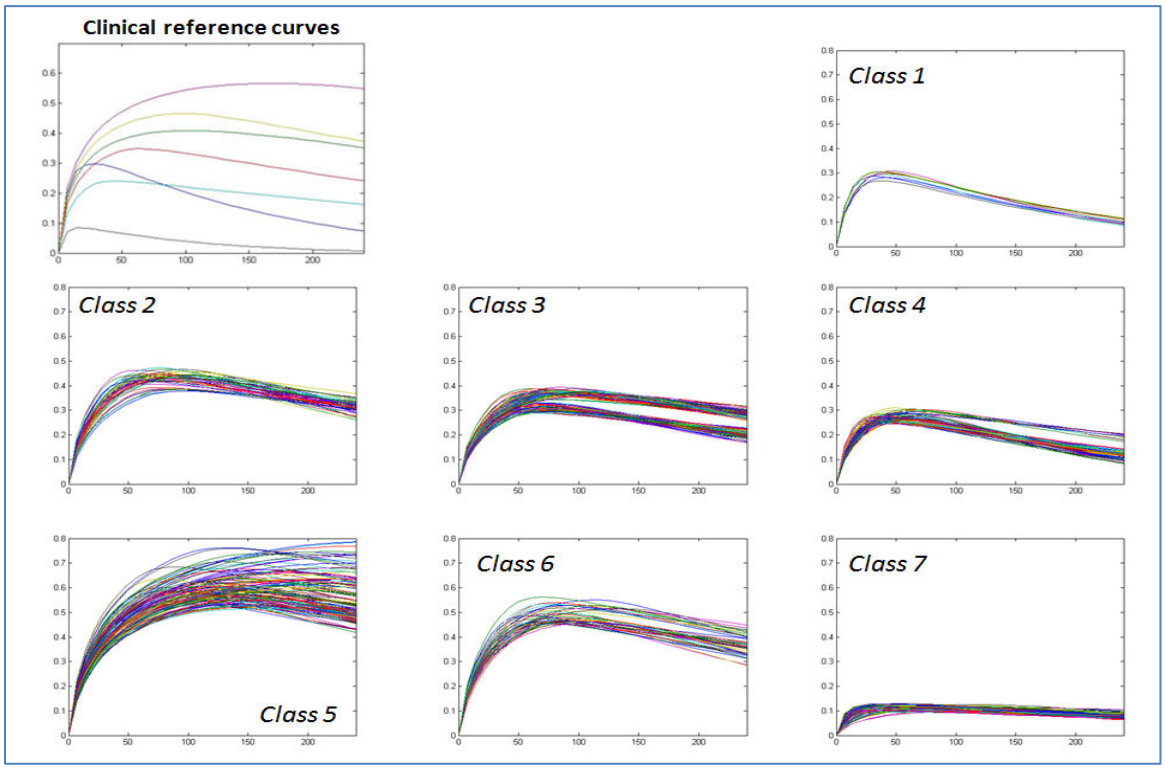


Figure 49. Data grouping of the experimental dataset based on the minimum New combined distance from the clinical reference curves.

In order to address both needs of shape and amplitude characteristics in clustering, we employ the proposed combined distance within the stabilized framework of the previous scenario. Thus, we perform repeated data partitioning through the k-means scheme and organize the meta-data (produced centroids) using the MSH approach. The Silhouette index for validation indicates an optimal number of classes again equal to 8. The final partition of the 497 samples is illustrated in Figure 48.

In Figure 49 is illustrated how the New combined distance reflects the cancer stages by measuring the minimum distance from the 7 clinical reference centroids. Although the New distance provides improved results in the case of reference curves centroids is approved that the clinical perception does not always match with the data structure.

The results demonstrate good potential of the proposed combined distance in preserving class characteristics. The 8 produced curves yield well separated and compact clusters. In particular, qualitative comparison of the clusters in Figure 48 with those in Figure 43 indicates more compact grouping in the former one. Furthermore, towards a more quantitative comparison, the confusion matrix of the new scheme with the combined distance metric appears in Figure 50. In comparison to Figure 25, the new partition shows better concordance with the true labels class 7, while it derives more balanced-split sub-classes (indexed 7 and 8). This self-evaluation process indicates an optimal number of classes equal to 8.

***Classification Results (derived labels per class)***

| *Reference labels per class* | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **count** |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| 3 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 71 |
| 4 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 71 |
| 5 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 71 |
| 6 | 0 | 0 | 0 | 0 | 0 | 64 | 0 | 7 | 71 |
| 7 | 0 | 0 | 0 | 0 | 0 | 7 | 28 | 36 | 71 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| count | 71 | 71 | 71 | 71 | 71 | 71 | 28 | 43 | |

Figure 50. Calculated confusion matrix along with the estimated IBSL curves regarding the implementation of scenario 1 for New combined distance.

## 5.4 Scenario 4: Bootstrap Clustering Approach Results

Consequently, from the previous two scenarios, we have established the use of bootstrap data resampling for stabilizing the clustering process, as well as the combined distance metric for preserving the shape and amplitude characteristics in clustering. In this last consideration, we aim at expanding the issues of exploratory self-organization to the case of big data using random sub-sampling in order to create multiple copies of the data encoding the various properties of the entire data set in smaller and more manageable pieces. In this form, we also test the generalization ability of bootstrap schemes by merging together the pieces of information carried by the bootstrap subsets. More specifically, we now test the stabilized scheme of scenario 1 with the combined distance of scenario 3 on bootstrap sub-samples taken in random from the big unlabeled set of 100000 samples. In each iteration a different bootstrap set (3000, 29) of centroids is randomly selected from a massive data clinical set (100000, 29). Following the process of stabilized clustering, a set of (2500, 29) centroids is formed. This new population is clustered with the MSH approach, leading to 8 principal classed from the Silhouette criterion, which are in Figure 51 (right side) along with the seven reference curves (left side) and the eight centroids derived from the smaller set of 497 samples in scenario 4 (center).

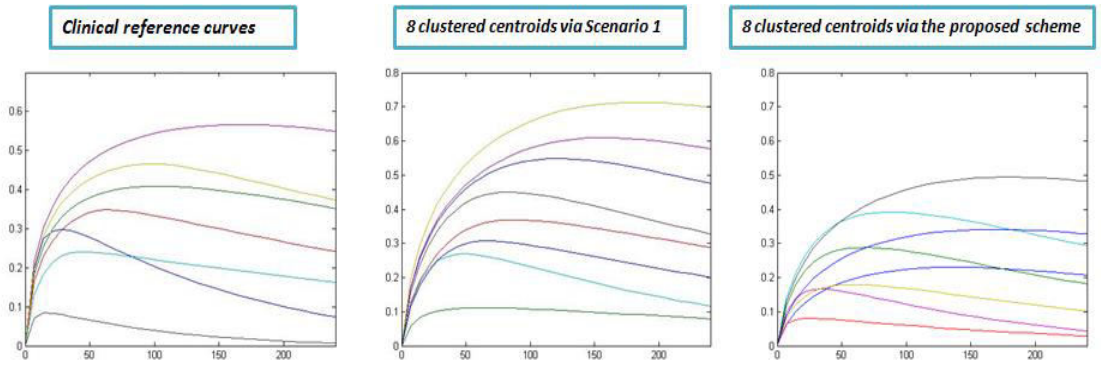


Figure 51. Clinical reference centroids (left), 8 derived centroids from scenario 1 (center) and estimated ones (right) representing the different tissue states via implementation of scenario 4.

The extracted clusters after the implementation of this scenario are depicted in Figure 52, demonstrating good separation of curves by shape and magnitude and good compactness of classes. At this point it is important to notice that a rather small size of the bootstrap population generated from only a small number of samples from the initial dataset is able to encode the primary data structure and reveal the shape of characteristic centroids. Attempts to increase the power of bootstrapping and tests on different dataset must be further contracted in order to establish the generalization ability, but the initial results from this study reflect good prospects of generalization to large populations from bootstrap subsamples.

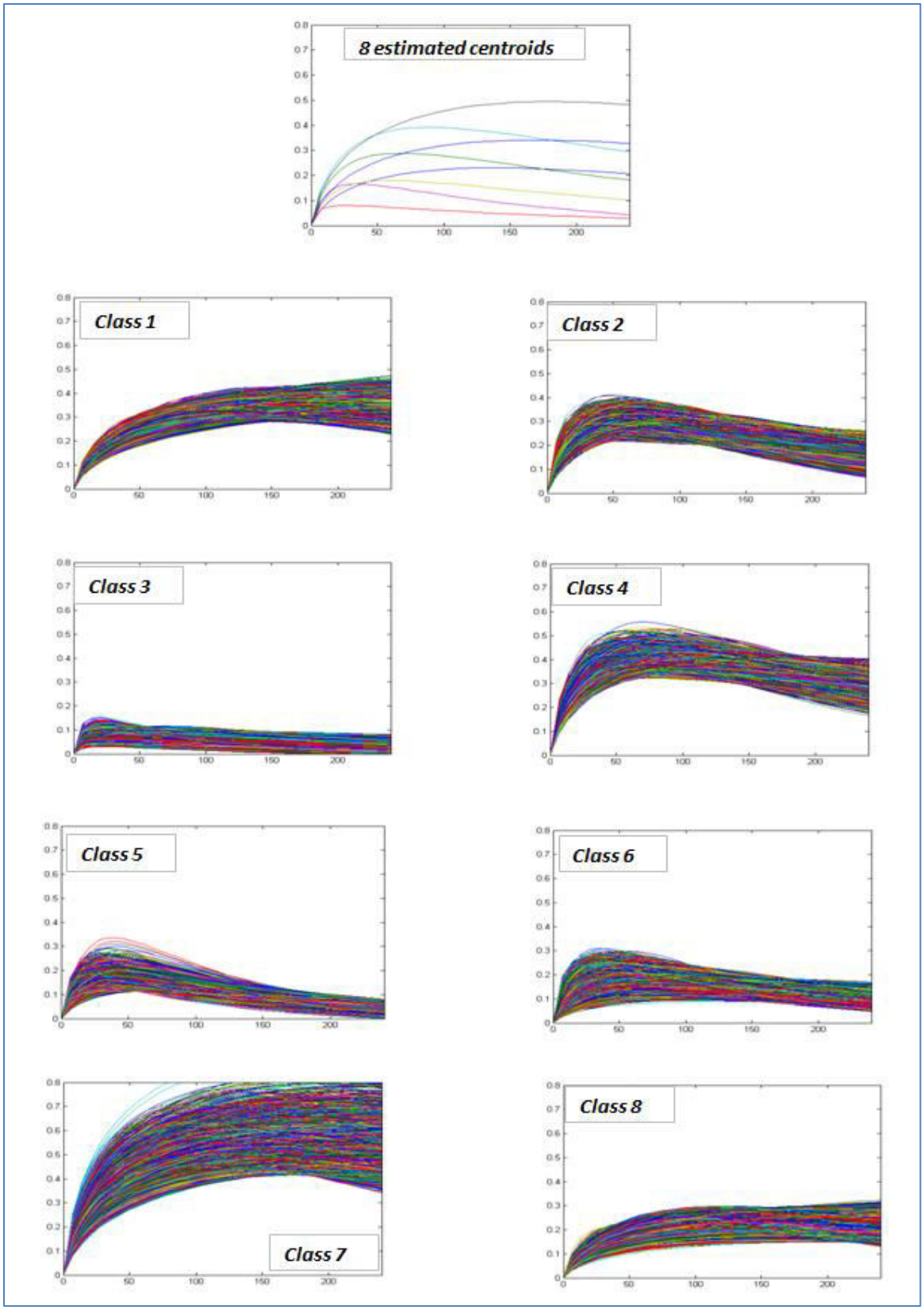


Figure 52. Qualitative representation of clusters based on the implementation of scenario 4, producing 8 classes.

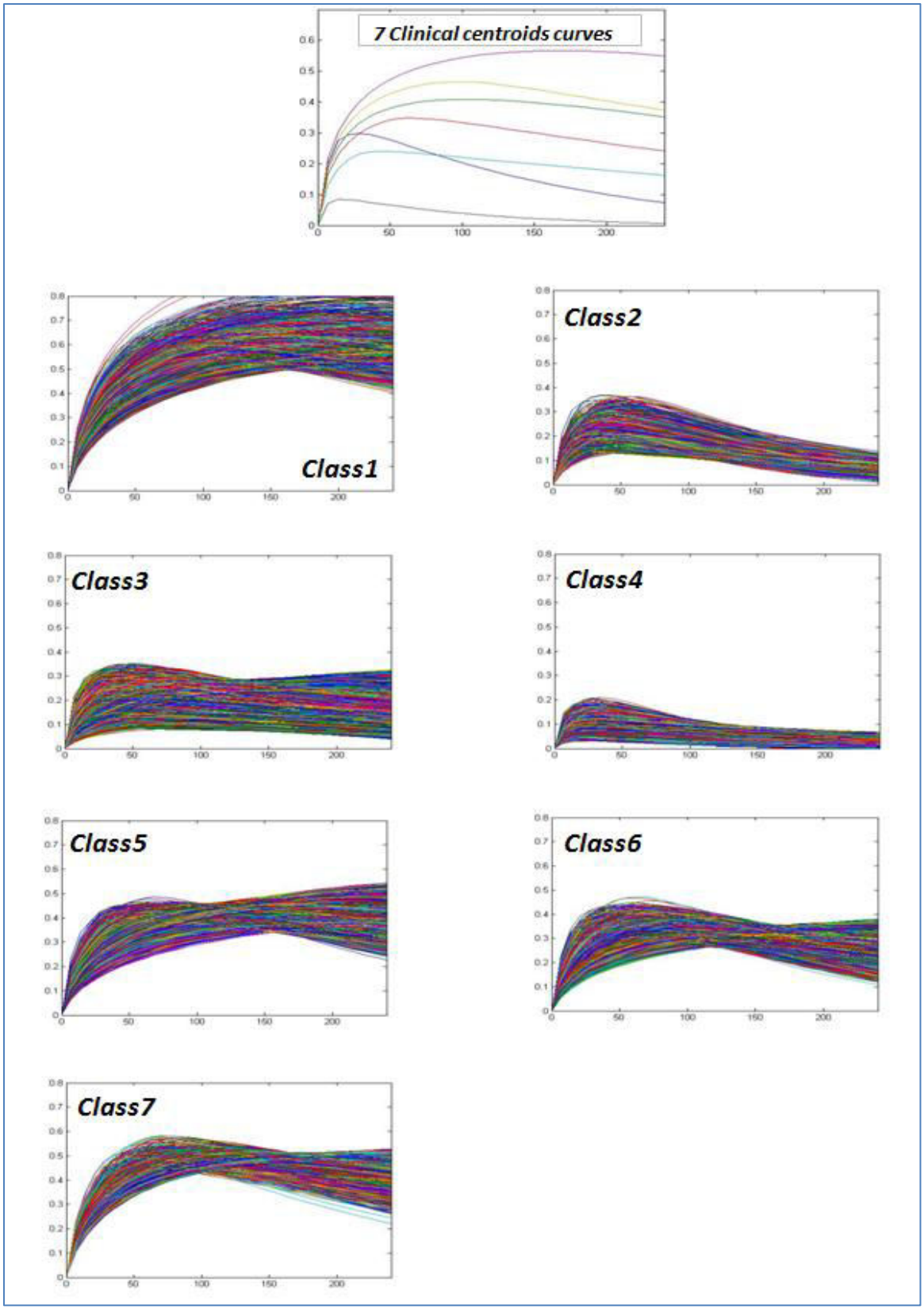


Figure 53. Clustering based on the minimum New combined distance from the reference clinical curves. Classes 5, 6, 7 are heavily mixed in shape and height.

In order to compare with similar clustering attempts we use the New combined distance in order to cluster the large set based on the minimum distance from the reference curves representing the clinical states of cervical tissue. The results are summarized in Figure 53, revealing that some classes (namely 5, 6 and 7) reflect severe mixing of curves different forms of curves and diverse amplitudes. Comparing the above clustering schemes, it is clear that the proposed scheme achieves good representation of clusters in both shape and amplitude terms.

By closer examination, we can verify that the mixed classes of Figure 53 have been decomposed in two classes In Figure 54. Furthermore, it can be noticed that the two hidden sub-classes have been revealed by our proposed scheme, expressing both pattern and magnitude differentiations. The consistent derivation of 8 classes by our self-organization schemes strongly suggests that additional clinical states exist and could be utilized in characterizing the grade of cervical neoplasia.

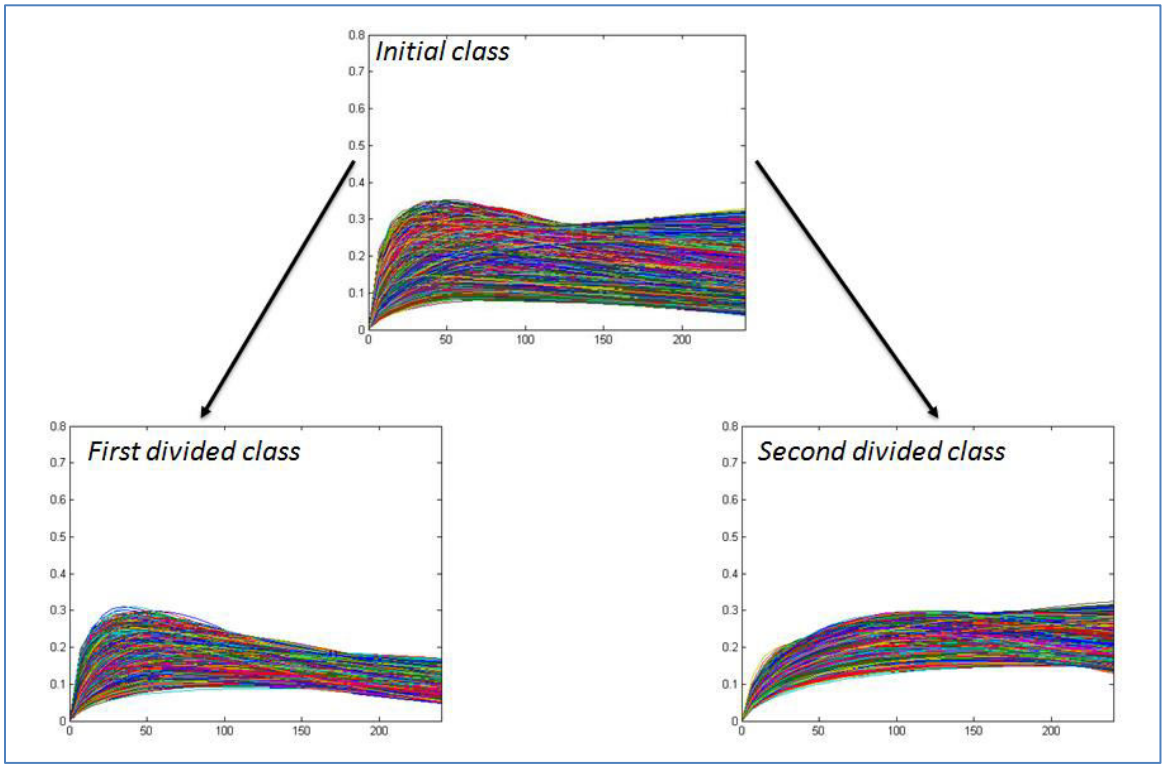


Figure 54. Qualitative illustration of two hidden sub0classes revealing the implementation of scenario 4.

It is interesting at this point to compare the outcome of this process stemming from the availability of big data with the clinical knowledge. The derived centroids from big data appear much closer to the clinical curves than the ones stemming from a much smaller sample (497 response curves). The exploitation of big data, which engage many response patterns and most likely encode all different types of trends in the progress of the AW phenomenon, enables the characterization of typical much closer to the ones reflecting the expert opinion. In addition, the analysis of big data reveals one additional pattern class that may potentially reflect a subclass of the pathology worth of clinical investigation.

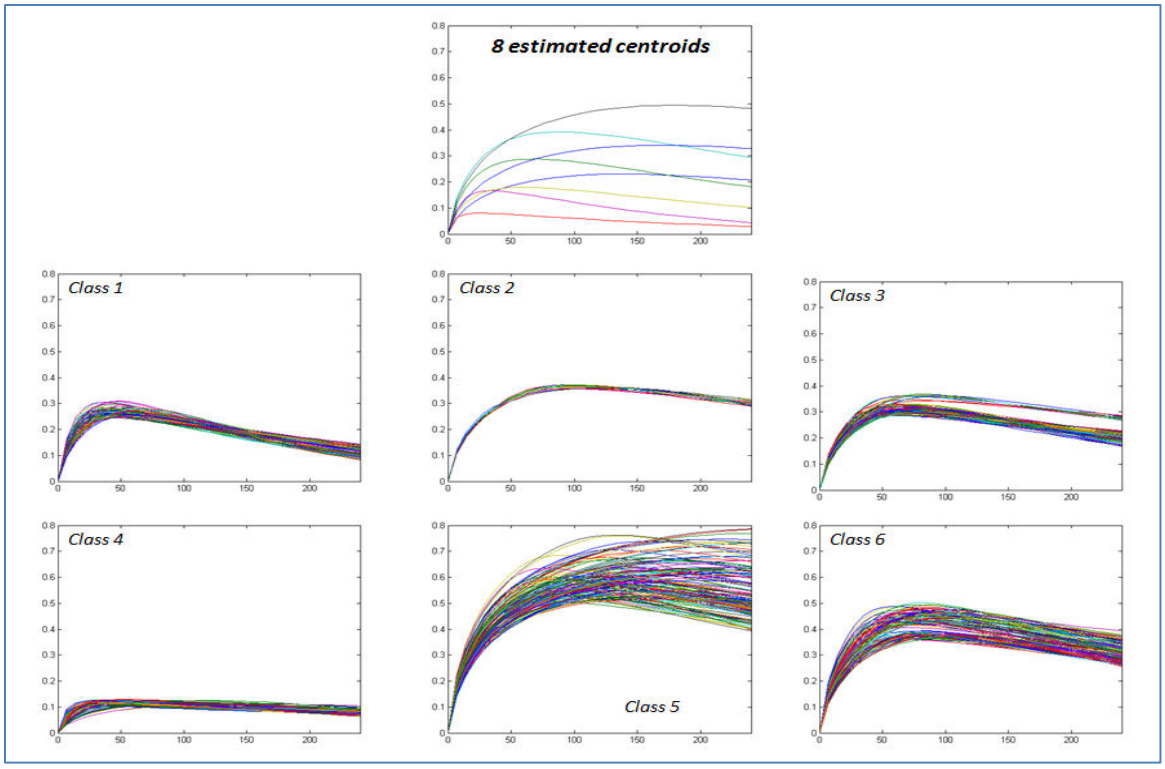


Figure 55. Data grouping of the experimental dataset based on the minimum New combined distance from the clinical reference curves

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 80 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 14 | 47 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 0 |
| count | 71 | 71 | 80 | 14 | 118 | 142 | 0 | 0 |

Figure 56. Quantitative results of experimental dataset based on the minimum New combined distance represented by confusion matrix.

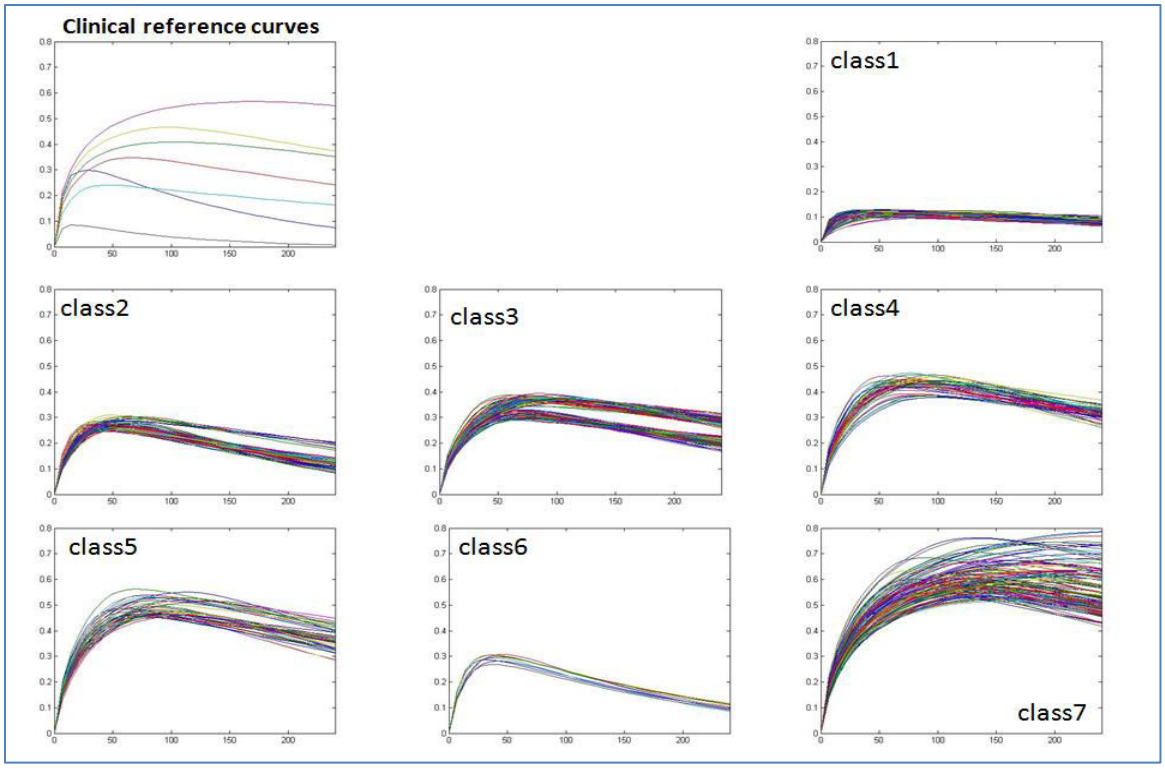


Figure 57. Data grouping of the experimental dataset based on the minimum Euclidean distance from the clinical reference curves.

***Classification Results (derived labels per class)***

| *Reference labels per class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 72 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 61 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 63 | 48 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 44 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 9 | 58 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 71 |
| count | 71 | 72 | 124 | 48 | 44 | 9 | 129 |

Figure 58. Quantitative results of experimental dataset based on the minimum Euclidean distance represented by confusion matrix

Using the small subset of responses (experimental dataset of 497 samples) cannot reflect the wealth of information in the staging of the pathology and the multiple form of the AW response within and across such stages. In fact, utilizing the centroids derived from big data to cluster the small subset of 497 curves based on the minimum new combined distance from the clinical reference curves results in mixed shape and amplitude forms, trend which is illustrated in Figure 55 where only 6 from the total 8 of extracted centroids have been connected to data samples. The same class diffusion effect under the previous clustering methodology is also observed on the confusion matrix, as illustrated in Figure 56, where it is clear that classes 7 and 8 originated from the big dataset are empty. The limited number of samples induces large interclass variances, which are statistically compensated in the multitude of the big dataset. The latter widespread distribution effect in data grouping is also observed utilizing the minimum Euclidean distance along with the seven reference centroids of clinical value. The corresponding outcome is qualitatively depicted in Figure 57 via the grouped IBSL curves and it is quantitatively presented via the resulting confusion matrix in Figure 58.

### ✓ Conclusions of Bootstrap Clustering Approach

In this work, we propose a novel framework for self-organization and grouping of datasets without prior knowledge on their statistical distribution and properties. The implemented methodology is based on the combined exploitation of clustering and bootstrapping approaches in order to reveal dominant groups from the entire population. The analysis is further enriched with an exploitation novel similarity metric that aims to capture both shape and amplitude differences from time-series data, so as to produce coherent and compact classes. For validation purposes, the proposed framework of analysis is applied on biomedical time-series data for cervical tissue evaluation and grading. More specifically, we consider an experimental dataset of 497 samples validated through actual biopsy and a large dataset of 100000 instances of unknown nature. We explore the potential of our approach to efficiently formulate, process and analyze representative clusters through a) repeated random sampling of the whole dataset in order to resolve stability issues and b) repeated random sub-sampling of big data in order to sample different properties of the distribution density of the dataset and, subsequent refinement of the cluster centroids based on a meta-clustering scheme applied on a large number of initial cluster centroids from the bootstrapping process. These methodological interventions are further supplemented with the use of a new distance metric, utilizing a weighted combination of known distance functions in order to reflect shape and magnitude resemblance on the clustered waveforms.

The major advantages of the presented methodology are its capability to automatically produce the final clusters without prior information on the formulation and nature of data, the effectiveness to search for and extract hidden characteristics and additional data structures within the examined population, as well as the significantly reduced complexity in association with the processing of big data, since large datasets are processed through representative subsets. An important side-issue of the proposed framework of analysis is the flexibility to interfere with information and processing tools at various levels of information abstraction, such

as in pre-processing (clustering of subsets from the dataset), intra-processing (using an alternative distance criterion) and meta-processing level (for self-organization of extracted cluster centers). In addition, its bootstrap formation aims at examining the data through different sub-spaces, offering the possibility to discover a wealth of information and stabilizing the clustering procedure from initialization effects.

The results of this study show that targeted alterations of simple clustering schemes, through bootstrap generation of sub exploitation-sets and automated meta-arrangement of extracted sub-centroids, can effectively organize and cluster large populations of time-series data, revealing a high potential for subclass discovery with significant contributions to early disease modeling, diagnosis and treatment. For the particular application considered, it would be challenging to investigate the importance of the additional eighth subclass of cervical tissue grading from the medical point of view. Moreover, the proposed approach could be generalized in the scientific field of data mining, serving as an efficient tool to detect hidden data structures and formations.

# 6. CONCLUSIONS & FUTURE WORK

Cervical cancer constitutes the most frequently diagnosed cancer type expressed in women worldwide. Apart from innovations in imaging techniques, an efficient methodology for extracting, processing and interpreting the relevant information from the available data is high importance. In general, the examined cases are represented by feature vectors reflecting the important properties of the tissue under examination or the characteristic structures from the imaging methodology. In this thesis, the data vector reflects the AW course over time for each pixel of the multispectral image in one specific wavelength range that best reflects the cellular deformations of cancerous tissue.

The goal of this study is to explore ways of organizing this data into meaningful classes acceptable by clinicians. Particularly, this thesis aims at characterizing each and every pixel of the recorded sequence of images over time, thus providing quantifiable measures for the state of the lesion and its borders. The clinician's knowledge is exploited in the beginning of the process by influencing the characteristic distributions of intended pathology classes, or at the end of clustering as a means of evaluating the quality and clinical value of automatically separated groups.

In algorithmic terms, this thesis attempts to overcome several limitations of self-organized clustering approaches pertaining to stabilization and generalization of the algorithms. Furthermore, it aims at combining expert knowledge with structural information from the data in order to make the data distributions more compatible with the clinically accepted response. We built our approach in four scenarios of incremental complexity on the assumptions and the strategic objectives aimed.

The first one considers the stabilization of the k-means algorithm and aims to make it robust over the initialization and the number of classes. In this sense, we propose a novel approach for automatically organizing data by fusing clustering and resampling approaches without any prior knowledge on their statistical distribution. Scenario 1 was applied to the training dataset of 497 characteristic IBSL curves. The results indicate the efficiency of this methodology to search for hidden information and detect statistical knowledge from unknown environments, giving rise to its utilization as an efficient tissue-characterization tool. Particularly, the proposed technique classifies the 497 sample into 8 final classes against the 7 ground-truth cervical cancer reference curves, indicating an extra clinical status of tissue in addition to the originally defined ones and shows good potential in discriminating smaller population of patterns within a possibly correlated environment.

However, this scheme reveals a difficulty in properly matching the estimated classes from a training dataset with biopsy results, as well as an inability to efficiently self-organize and classify the testing dataset of 100000 cases into well discriminated and meaningful groups of curves. Thus, a different more intuitive and advanced approach should be considered. This technique needs generalization in order to apply to completely unknown environments with large data set. Furthermore, after the examination of two distance metrics, we can verify that the Euclidean distance classifies by the magnitude but looses the shape, while the Squared Cosine distance

gives priority to the shape. Consequently, the necessity of a new distance metric, that will combine the above important traits, becomes essential. Finally, we also address the need to utilize the clinical knowledge (or the clinically accepted classes with their centers).

In order to address these issues we next develop the second scenario, where we propose a novel algorithmic scheme for self-organizing data, adopting a recursive k-means mode utilizing available domain knowledge and data-hidden information with the aim to efficiently classify the training set of 497 samples. The stability of bootstrapped k-means scheme through multiple initializations is examined in this scenario. Furthermore, we combine both clinical and data information by building an incremental scheme where the centers initialized by expert knowledge are updated and incrementally corrected via the data distributions. The proposed recursive k-means approach provides an efficient classification of the training set. More specifically the resulting clustered centroids are clearly distinguished one from the other, forming compact and well separated clusters.

The results indicate certain advantages of the proposed recursive k-means case initialized by reference seeds against the classic k-means approach. However, it is important that this method can search for hidden information from completely unknown environments. Finally, even though the above schemes achieve better agreement on the structure conveyed by the data and the expert-defined shapes, there is still certain mismatch that questions the validity of metrics used for comparing the time-series data curves.

Thus, we examine in the third scenario several combinations of distance measures and propose a novel metric stemming from such a joint consideration, which places importance in both the magnitude and shape of data. The results regarding scenario 3, demonstrate good potential of the proposed combined distance metric in preserving class characteristics.

Finally, in the fourth scenario we examine the generalization potential of bootstrapping techniques, considering the clustering of bid data in smaller subsamples and the combination of intermediate results towards a more detailed and accurate characterization of the entire dataset. We examine the validity of these schemes on the testing data set of 100000 samples. The results reveal the good capability and potential of the proposed technique to automatically organize compact clusters without prior information on the formulation and nature of the data. Moreover, the proposed approach is efficient to search for and extract hidden characteristics and additional data strictures within the examined population. An added advantage of this approach is the significantly reduced complexity in association with the processing of big data, since large datasets are processed through representative subsets.

An important methodological contribution of the thesis is the flexibility to interfere information and apply processing tools at various levels of information abstraction, in pre-processing (clustering of subsets from the dataset), intra-processing (using an alternative distance criterion) and meta-processing level (for self-organization of extracted cluster centers). In addition, its bootstrap formation attains stabilization of the clustering procedure from initialization effects, but also aims at examining the data through different sub-spaces, thus offering the possibility to discover a wealth of data-hidden information.

The results of this study show that targeted alterations of simple clustering schemes, through bootstrap generation of sub exploitable sub-sets and automated meta-arrangement of

extracted sub-centroids, can effectively organize and cluster large populations of time-series data, implying a high potential for subclass discovery with significant contributions to early disease modeling, diagnosis and treatment. Moreover, the proposed approach could be generalized in the scientific field of data mining, serving as an efficient tool to detect hidden data structures and formations.

Future work in this direction includes the improvement of the selection procedure to determine the optimal size of the bootstrap subsets, the exploitation and testing of different datasets from various application areas, the examination of confidence intervals on the statistics utilized, as well as the adoption of alternative validity metrics to assess the robustness and significance of the number and composition of extracted classes.

Another challenging issue to deal with under this algorithmic development framework is the exploration of the sequential distance metric, which we presented in scenario 3. Despite the ability of this approach to include and preserve the properties of both distances, squared Euclidean distance and Squared Cosine distance, further investigation of the appropriate merging criteria is needed in order to validate the form of the final centroids.

Furthermore, a topic of high importance regarding the results of this thesis concerns the evaluation of the final classes. It is worth mentioning, that there are two dominant modes within the characteristic IBSL curves, which are used in this thesis. The first mode increases gradually and maintains the high level for a while. The second curve climbs moderately, reaching a peak and after some seconds declines sharply. The clinical importance and the tissue-related origin of these two types of curves should be investigated from a medical perspective. Furthermore, the proposed approach reveals an additional sub-class. Consequently, it would be challenging to investigate the importance of the additional eighth subclass of cervical-tissue grading from the clinical point of view. Overall, the combination of clinical knowledge with data-hidden information, as well as the evaluation of subclasses revealed by the data structure could lead to very interesting developments.