FCM PARAMETER ESTIMATION METHODS: APPLICATION TO INFRARED SPECTRAL HISTOLOGY OF HUMAN SKIN CANCERS

T. Happillon[1], D. Sebiskveradze[1], V. Vrabie[2], O. Piot[1], P. Jeannesson[1], M. Manfait[1], C. Gobinet[1]

[1]Unité MéDIAN, CNRS FRE 3481 MEDyC, UFR de Pharmacie, Université de Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims Cedex, France
[2]CReSTIC, Université de Reims Champagne-Ardenne, Moulin de la Housse, BP 1039, 51697 Reims Cedex, France

ABSTRACT

Spectral histology of cancer can be achieved thanks to the analysis of infrared (IR) hyperspectral images. The Fuzzy C-Means (FCM) clustering is particularly well adapted since each object is attributed to all clusters with different membership values. Applied on IR hyperspectral images of human skin cancers, it can highlight fine transitions between tumor and surrounding tissues and/or tumor heterogeneities. However, to provide a biomedically interpretable clustering, the relevant values of the two FCM parameters, i.e. the number of clusters and the fuzziness parameter, must be judiciously selected for each analyzed tissue. In this paper, the performance of some classical cluster validity indices and *m*-rules previously presented in the literature are evaluated. A new heuristic method based on the redundancy of FCM clusters is also proposed. We show that our method highly improves the clustering quality and computational time when applied on IR images of human skin tumors.

Index Terms— Infrared spectral imaging, skin cancer, Fuzzy C-Means, validity indices, *m*-rules, spectral histology

1. INTRODUCTION

A challenge in oncology is to early diagnose tumors in order to increase the therapeutic effects of medical treatments and improve the life expectancy of patients. Traditionally, the diagnosis is realized by conventional histology which is based on the morphological analysis of biopsy slices stained with Hematoxylin and Eosin (HE). Since few years, novel and non-destructive biophotonic approaches are developed in order to achieve spectral histology as a complement to conventional histology.

Infrared (IR) spectroscopy has been shown to be a potential candidate to achieve spectral histology for cancer diagnosis. This technique is based on the interaction of an incident IR light beam with a point on the analyzed sample. When the energy of an incident photon is equal to the energy of a vibrational mode of the sample, this photon is absorbed by the sample, and a decrease of the transmitted light intensity is recorded. These intensity losses are recorded at different wavelengths (or wavenumbers) to give the absorbance spectrum of the acquisition point. A spectrum is thus a real molecular fingerprint giving information about the structure and metabolism of biomolecules at the acquisition point. IR spectroscopy has already been used to discriminate between different kind of tumors and also between healthy and tumoral tissues [1].

By repeating this operation with a scanning of the sample surface, a hyperspectral dataset is recorded. This huge amount of data requires the use of clustering methods to achieve spectral histology [2, 3], i.e. for the histopathological recognition of the sample structures, especially to highlight some structures invisible for the pathologist on HE histological images. Hierarchical analysis and K-Means (KM) clustering have been shown as very efficient to retrieve the analyzed tissue structures and to localize tumors [3, 4]. However, each spectrum belongs to only one cluster with these techniques.

Fuzzy C-Means (FCM) [5] seems to be better adapted since each spectrum can be attributed to all clusters thanks to membership values ranging from 0 to 1. This property is biologically relevant since transitional zones between tissue structures or even heterogeneities can be highlighted, while it is not always the case on the HE images.

FCM is controlled by two parameters, i.e. the number of clusters C and the fuzziness parameter m, that must be fixed by the user, and from which will depend the quality of the clustering results. When applied to IR spectroscopy, C is usually empirically chosen, whatever the clustering method used [3, 4], based on biological knowledge on the sample composition. On the contrary, m is classically fixed to 2 [3].

Some solutions have been developed to answer this parameterization problem. First, the automatic selection of C has been addressed by the so-called validity indices which optimize a criterion [6]. Second, *m*-rules estimate m by the optimization of a criterion [7-9] or by using empirical rules [10]. However, validity indices and *m*-rules have been developed independently from each other. Validity indices assume that m=2, while some applications on real world dataset show that this choice can lead to misclutering since it depends on the structure of the data. Moreover, some *m*-rules suppose that C must be *a priori* known.

The development of methods able to simultaneously determine *C* and *m* is thus crucial. Recently, a new method, named FCM-Redundant Based Algorithm (FCM-RBA) has been proposed to solve this problem [11]. This method is based on the removal of redundant clusters. The first aim of this paper is to propose a faster version of this heuristic method. The second aim is to compare the efficiency of classical validity indices, *m*-rules and our method on IR images of human skin tumors.

2. FUZZY CLUSTERING TECHNIQUE

FCM [5] is a famous clustering algorithm which can be seen as a generalization of the well-known KM algorithm by introducing a fuzziness parameter *m*. This parameter must be higher than 1. Indeed, a value of *m* close to 1 makes FCM equivalent to the crisp KM clustering. On the contrary, as *m* increases, the fuzziness of the results increases as well. If *m* tends to infinity, the estimated clusters are identical.

With FCM, each IR spectrum can belong to every cluster with a membership coefficient ranging between 0 and 1. The only constraint is that the sum of these coefficients must be equal to 1 for each spectrum.

Intermediate spectra could thus be highlighted as their membership coefficients are equally distributed between several clusters. This kind of dataset clustering emphasizes a new type of information like transitional zones between two or more clusters. Such transitions are very informative here since they could match to a tumor progression into a surrounding healthy tissue and/or tumor heterogeneity.

In order to highlight those zones, it is necessary to determine the optimal values of *C* and *m*. That's why cluster validity indices and *m*-rules have been developed.

3. CLUSTER VALIDITY INDICES

Cluster validity indices estimate the optimal value of *C* by looking for the value of *C* that optimizes a criterion [6]. To construct these validity indices, several FCM with different values of *C* must be computed, being very time-consuming.

The first validity indices only consider the membership coefficients estimated by FCM [6]. We can cite the partition coefficient (V_{PC}) based on the sharing of objects between clusters, the partition entropy (V_{PE}) based on the fuzziness of the estimated partition, the modified partition coefficient (V_{MPC}) which is a transformation of V_{PC}, the Windham proportion exponent (V_{WPE}) based on the overlapping between clusters, the Kim-Kim-Lee-Lee coefficient (V_{KKLL}) based on the sharing between clusters, and the fuzzy partition measure (V_P) based on compactness and separation notions [6].

To overcome the observed sensitivity of these indices to *m* and their monotonic behavior in function of *C*, new validity indices were developed using membership coefficients and centroids [6]. A majority of these indices are constructed by melting two approaches, inter-cluster distances, which should be as biggest as possible, and the intra-cluster distances, which are preferred when they are as smallest as possible. The most efficient indices in the scientific literature are the following ones [6]: the Fukuyama-Sugeno index (V_{FS}), the Xie-Beni index (V_{XB}), the Kwon index (V_{Kw}), the Tang index (V_T), the partition coefficient and exponential separation (V_{PCAES}), the validity index of Tsekouras and Sarimveis (V_{VI}), the compose within and between scattering (V_{CWBS}), the Wang-Sun-Jiang index (V_{WSJ}), the Pakhira-Bandyopadhyay-Maulik index (V_{PBMF}), the separation and total compactness index (V_{STC}), the fuzzy separation and compactness index (V_{FSC}). Other indices are based on the hypervolume and the density (the fuzzy hypervolume index (V_{FHV}), the partition density index (V_{PD}) and the average partition density (V_{APD})), the granularity-dissimilarity measure (V_{GD}), the ratio of the overlapping to the separation (V_{OS}) or an optimality test (V_{OT}). Due to the lack of place, the interested reader can refer to [6] and references therein for a detailed presentation of these validity indices.

4. M-RULES

The aim of *m*-rules is to find the value of *m* which optimizes a criterion. *m*-rules can be decomposed into two groups. The *m*-rules belonging to the first group only use the data to assess the optimal value of *m*. Among these methods can be cited the Dembélé-Kastner empirical rule (R_{DK}) which is based on the computation of the variation coefficient, or the theoretical β rule (R_β) which computes upper bound for *m*.

The second group is composed of rules based on the results of the FCM algorithm, such as the fuzzy decision optimization rule (R_{FDO}) which maximizes the intersection between a fuzzy goal and a fuzzy constraint, or the Okeke-Karnieli rule (R_{OK}) which minimizes the reconstruction norm of the dataset with the FCM results.

5. FCM-RBA

In [11], a new heuristic method, named FCM-RBA, determining automatically the best values of *C* and *m* was proposed. Its main idea is to decrease the number of clusters if at least two clusters are similar for a given *m*.

The similarity here is measured by the Pearson correlation coefficient between the membership values. *T* is a new introduced parameter which represents a threshold for the correlation coefficients. It takes its values between 40% and 95%, with a step of 5%. If there is redundancy between two clusters, i.e. their correlation coefficient is superior to *T*, then one of them becomes useless.

In this article, FCM-RBA is modified in order to improve the computational time. First, an interruption inside the FCM algorithm itself occurs when the current results, during successive iterations, present redundancy between at least 2 clusters. Second, FCM results being independent on *T*, if, for a previous *T* value, a FCM clustering has already been computed for a given setting of *C* and *m*, then it is not necessary to

compute it again.

This redundant method proceeds as follows:

1. Initialization of parameters. C is empirically initialized at 15, m at 1.1 and T at 0.4.
2. Calculation of FCM (C, m, T).
3. Interruption of FCM, allowing a gain in time complexity, if redundant clusters are detected. C is revalued as follows: $C = C–1$, and FCM (C, m, T) is recomputed by going to step 2.
4. If no redundancy is detected in step 3, the best value of C is recorded into $C_{best}(m, T)$.
5. All those calculations are repeated from the second step with a new value of m, $m = m + 0.1$ and initialization of C as $C = C_{best}(m-0.1, T) + 2$, if the floor of 15 is respected, otherwise $C=15$. It is interesting to mention that this initialization of C induces a considerable gain in time of execution.
6. When all the values of m are gone over, all those calculations are run again with a new value of T as $T = T + 0.05$, till T is equal to 0.95.

For each T, a curve $C_{best}(m, T)$ is thus estimated. The interpretation of these curves to find the best m and C is left to section 6.4.

6. RESULTS AND DISCUSSION

6.1. Sample description

The previously described methods have been applied on an IR hyperspectral image acquired on a basal cell carcinoma (BCC) section. The image is composed of 104x61 pixels. In each pixel, a spectrum composed of 451 wavenumbers uniformly distributed between 900 and 1800 cm^{-1} was recorded. The spectra are considered as the objects to be clustered.

Figure 1 shows the HE image of the section adjacent to the one analyzed by IR imaging and annotated by a pathologist who is able to distinguish 4 different structures: stratum corneum (1), epidermis (2), dermis (3) and tumor nests (4). The dermis is known to be heterogeneous and can thus be decomposed into several clusters [4, 11].

6.2. Validity indices and m=2

We run validity indices with m=2 as traditionally described in literature. The estimated optimal numbers of clusters C_{opt} are presented in the second column of table 1. The majority of the validity indices estimates C_{opt}=2. Figure 2 shows the clusters estimated by FCM with such parameters. Cluster 2 represents mainly the dermis, while the epidermis and the tumor nests are merged in cluster 1. For 3 clusters such as estimated by V_{VI}, the results are the same except that the dermis is divided into two clusters (data not shown). The other validity indices are optimized for $C_{opt}\geq 11$ which induces a large number of redundant clusters (data not shown). FCM run with 11 clusters, such as estimated by V_T, leads to 4 identical clusters representing both tumor nests and epidermis (cluster 1, figure 3), 5 identical clusters of dermis (cluster 5, figure 3), and two supplementary different clusters of other dermis structures (clusters 10 and 11, figure 3). A m-value equal to 2 thus induces too much fuzziness for this spectral dataset in the FCM model and tumor nests cannot be separated from the epidermis. It is thus necessary to adapt the m-value to the studied dataset thanks to the use of the m-rules.

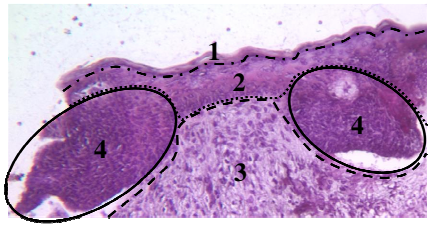

Figure 1: HE histological image of the BCC sample.

Validity indices	C_{opt} for m=2	C_{opt} for m=1.1	C_{opt} for m=1.3	C_{opt} for m=1.9	C_{opt} for m=4.4
V_{PC}	2	2	2	2	2
V_{PE}	2	2	2	2	2
V_{WPE}	2	2	2	2	2
V_{MPC}	2	4	5	2	3
V_{KKLL}	2	10	5	2	2
V_{P}	2	4	5	2	3
V_{FS}	15	14	13	15	15
V_{XB}	2	4	3	2	10
V_{Kw}	2	4	3	2	3
V_{T}	11	2	2	3	2
V_{FSC}	2	2	2	2	2
V_{FHV}	2	2	2	2	2
V_{APD}	2	2	2	2	2
V_{PD}	2	2	3	2	2
V_{PCAES}	2	2	2	2	8
V_{VI}	3	4	5	3	15
V_{CWBS}	14	4	5	11	6
V_{WSJ}	14	4	5	13	10
V_{PBMF}	15	2	2	15	15
V_{STC}	2	3	3	2	3
V_{GD}	2	4	3	2	3
V_{OT}	2	2	2	2	2
V_{OS}	2	5	5	2	11

Table 1: Optimal number of clusters C_{opt} estimated by the validity indices for different values of the fuzziness parameter m.

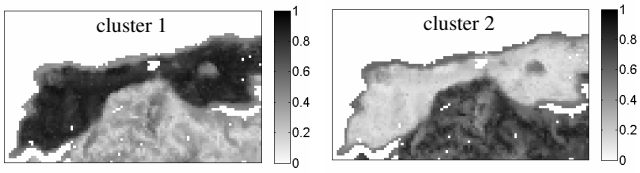

Figure 2: Membership values estimated by FCM for C=2 and m=2.

m-rules	R_{DK}	R_{OK}	R_{FDO}	R_{β}
m_{opt} for C=5	1.1	1.3	1.9	4.4

Table 2: Optimal m-value m_{opt} estimated by the m-rules.

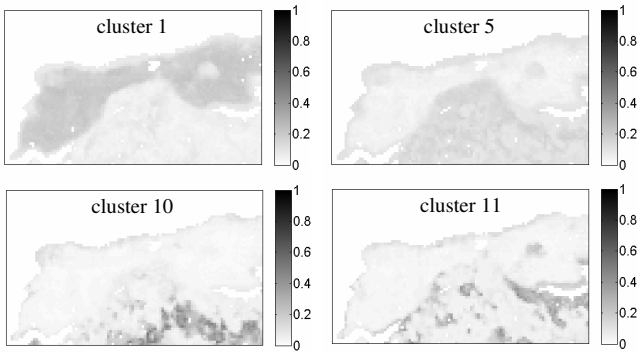

Figure 3: Membership values estimated by FCM for C=11 and m=2.

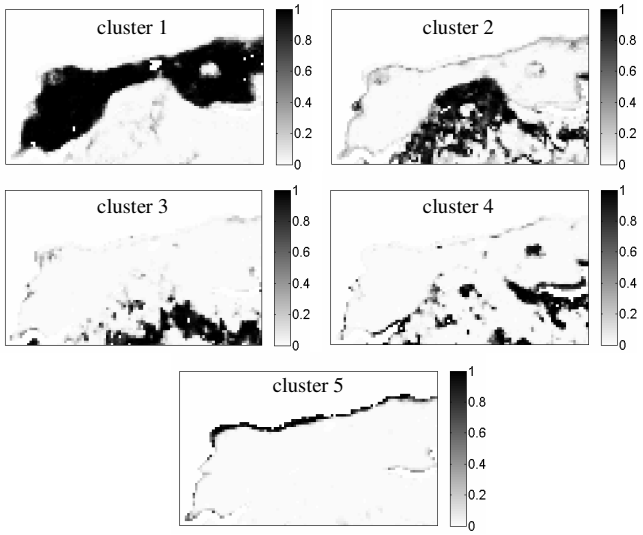

Figure 4: Membership values estimated by FCM for C=5 and m=1.3.

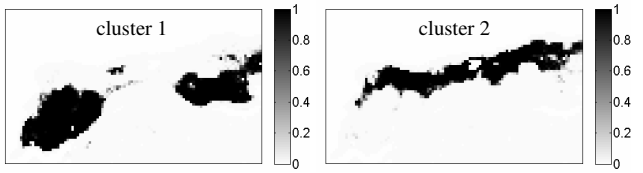

Figure 5: Membership values estimated by FCM for C=10 and m=1.1. Only the first 2 clusters are shown.

6.3. Validity indices and *m*-rules

R_{FDO} and R_{OK} being dependent on the number of clusters, we applied these rules by assessing that C=5 such as estimated by FCM-RBA and validated by a pathologist in another study [11]. The optimal m-values m_{opt} estimated by the m-rules are presented in table 2.

The optimal number of clusters C_{opt} was estimated by each validity index for the optimal m-value m_{opt} assessed by each m-rule. The results are pooled in columns 3 to 6 of table 1.

For m_{opt}=4.4 as estimated by R_β, it is clear that the results will be fuzziest than those estimated with m=2. For m_{opt}=1.9 as estimated by R_{FDO}, the same results as m=2 are obtained.

For m_{opt}=1.1 as estimated by R_{DK} or m_{opt}=1.3 as estimated by R_{OK}, the majority of the validity indices is optimized for $C_{opt}\leq$5. Figure 4 shows the FCM results for C_{opt}=5 and m=1.3 as estimated by R_{OK}. Tumor nests and epidermis are still merged (cluster 1). Clusters 2, 3 and 4 represent different dermis structures. Cluster 5 is the stratum corneum.

The other validity indices estimated $C_{opt}\geq$10. For such number of clusters, the tumor nests and the epidermis are isolated on different clusters as can be seen on figure 5 (clusters 1 and 2 respectively). However, the dermis is divided on 5 clusters, the stratum corneum on 2 clusters, and one cluster is composed of pixels at the interfaces stratum corneum/epidermis-tumor and epidermis-tumor/dermis (data not shown).

Only V_{KKLL} combined with R_{DK}, and V_{FS} combined with R_{DK} or R_{OK} are able to retrieve clusters associated to the tumor nests. However, the number of clusters necessary to isolate the tumor nests is high, and the optimal m-values estimated by R_{DK} and R_{OK} are near from 1, which means that the obtained results are comparable to those obtained with KM in another study [4] on the same sample with an empirical estimation of the number of clusters.

These results thus suggest that m-rules and validity indices are not adapted to this kind of data.

6.4. FCM-RBA

The curves $C_{best}(m,T)$ for the 12 different values of T obtained by FCM-RBA are presented on Figure 6. These curves rapidly decrease till a break which has been interpreted as a stabilization of the FCM solution. Thus, the best values of the parameters C and m are considered as being pointed out by the first break of each curve. Indeed, 7 curves present their curve break for C=5 and m=1.6, and 5 for C=5 and m=1.5 or 1.7. It must be known that making a clustering with a variation of 0.1 on the m-value doesn't change much the FCM results. Consequently, the best parameters are chosen as C=5 and m=1.6. Figure 7 presents the gray scale images obtained by FCM run with those parameters. Estimated clusters represent all the different tissue structures of the biopsy which were observed by the pathologist: the dermis, divided into three clusters because of its heterogeneity, the epidermis and the tumor nests. The interested reader can refer to [4] in order to access to the KM results obtained on this biopsy.

6.5. Discussion

The majority of validity indices used in this study is initialization dependent. On the contrary, our new method is more robust against initialization because the determination of the relevant FCM parameters is based on several curve breaks (one for each threshold value) and not on a unique maximum or minimum value such as the validity indices.

Furthermore, our method has been developed in order to simultaneously estimate m and C, while validity indices

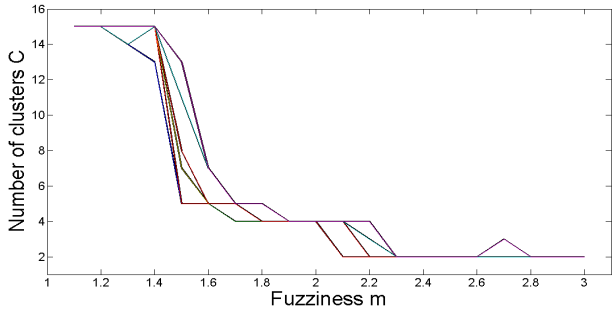

Figure 6: Curves $C_{best}(m,T)$ for the 12 different values of T estimated by FCM-RBA allowing identification of C_{opt} and m_{opt}.

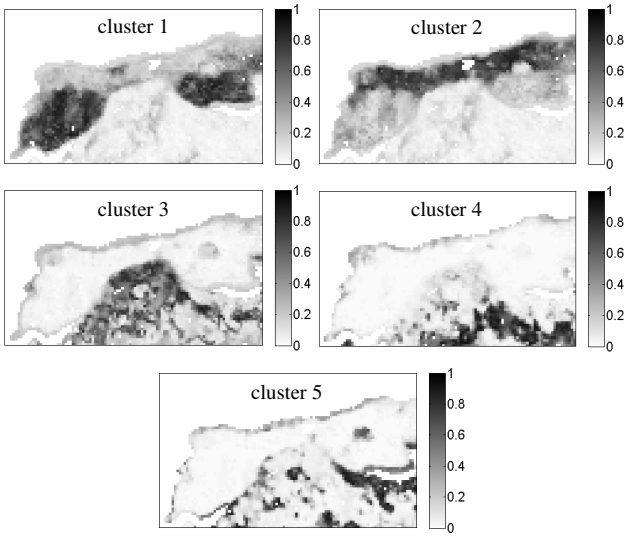

Figure 7: Membership values estimated by FCM for the parameters values given by FCM-RBA.

and *m*-rules have been developed ones independently from the others.

The main improvement is that, compared to the algorithm proposed in [11], the algorithm presented in this article is 8 times faster. This is due to a considerable diminution of the number of computed FCM. Indeed, many irrelevant settings of FCM are avoided, and interruption of FCM is imposed when redundant clusters are estimated.

These results were also validated on a group of 15 other biopsies representing several types of skin tumors (results not shown here due to the lack of place).

7. CONCLUSION

FCM is a clustering method that is well adapted to the analysis of IR spectral images acquired on human skin tumors. The FCM parameters, i.e. the number of clusters and the fuzziness parameter, must be carefully chosen. Several validity indices and *m*-rules have been proposed in the literature to overcome this selection problem. They have been compared in this paper to a new heuristic method based on the redundancy of clusters estimated by FCM. Our proposed method, named FCM-RBA, leads to better results in term of quality than those obtained with the validity indices combined to the *m*-rules. Indeed, this method is qualitatively more efficient as shown on a real dataset. Each cluster represents a unique tissue structure composing the biopsy, enabling an easy localization and characterization of the tumor. Moreover, the heterogeneity and the possible transitions between tissue structures could be highlighted, which cannot be realized with KM algorithm or, in most of cases, with a classical HE image analyzed by a pathologist. That's the reason why this new kind of information is of very important interest from a biomedical point of view.

8. REFERENCES

[1] R.K. Dukor, *Handbook of Vibrational Spectroscopy*, John Wiley and Sons Ltd., New York, 2001.

[2] B. Bird, M. Miljkovic, M. J. Romeo, J. Smith, N. Stone, M. W. George, and M. Diem, "Infrared microspectral imaging: distinction of tissue types in axillary lymph node histology," *BMC Clinical Pathology*, vol. 8, n°8, 2008.

[3] P. Lasch, W. Haensch, D. Naumann, and M. Diem, "Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis," *Biochimica et Biophysica Acta*, vol. 1688, pp. 176-186, 2004.

[4] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M.Manfait, "Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies," *Analyst*, vol. 133, pp. 197-205, 2008.

[5] J.C. Bezdeck, *Pattern recognition with fuzzy objectives function algorithms*, Plenum Press Ed, New York, 1981.

[6] W. Wang, and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets and Systems*, vol.158, pp. 2095-2117, 2007.

[7] J. Yu, Q. Cheng, and H. Huang, "Analysis of the weighting exponent in the FCM," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 34, pp. 634–639, 2004.

[8] H. Choe, and J. B. Jordan, "On the optimal choice of parameters in a Fuzzy C-Means algorithm," *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 349–354, 1992.

[9] F. Okeke, and A. Karnieli, "Linear mixture model approach for selecting fuzzy exponent value in Fuzzy C-Means algorithm," *Ecological Informatics*, vol. 1, pp. 117–124, 2006.

[10] D. Dembélé, and P. Kastner, "Fuzzy C-Means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973–980, 2003.

[11] D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson, and O. Piot, "Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections," *Laboratory Investigation*, vol. 91, pp. 799-811, 2011.