# Indices de formes et de textures Application au classement de noyaux de cellules

**Guillaume Thibault — Bernard Fertil — Jean Sequeira — Jean-Luc Mari**

*Université de la Méditerranée (Aix-Marseille 2), LSIS*
*Case 925 - 163 avenue de Luminy 13288 Marseille cedex 9*
*Guillaume.Thibault@univmed.fr*

RÉSUMÉ. *Cet article présente une étude sur le diagnostic de noyaux de cellules sanguines provenant de patients atteints de Progéria. Les noyaux sont caractérisés à l'aide de différentes méthodes d'analyse de forme et de texture, puis classés par des techniques d'apprentissage.*
*Les mesures extraites des noyaux servent aux calculs d'un ensemble de valeurs nommées des « indices ». Les indices sont des fonctions apportant des valeurs scalaires ou discrètes qui nous renseignent sur les caractéristiques géométriques, morphologiques et texturelles des noyaux. Toutes ces informations numériques permettent de construire un modèle de description des noyaux qui est ensuite utilisé pour les classer.*

ABSTRACT. *This paper present a study on the diagnoses of blood cell nuclei from patients affected by the Progeria disease. Nuclei are characterized with various methods that analyze them shape and texture, and subsequently classify them by learning techniques.*
*These methods extract "measures" from nuclei, which are used for the computation of a set of values named "indexes". These indexes are function bringing up real or discrete values that give information about geometrical, morphological and textural features of nuclei. A description model of nuclei is built based on this data, that is used to classify nuclei.*

MOTS-CLÉS : *Indices de formes et de textures, caractérisation de formes et de textures, matrice de surface de zones, classement.*

KEYWORDS: *Shape and texture indexes, shape and texture characterization, gray level size zone matrix, classification.*

RSTI - ISI – 15/2010. Traitement de données complexes, pages 73 à 97

## 1. Introduction et contexte

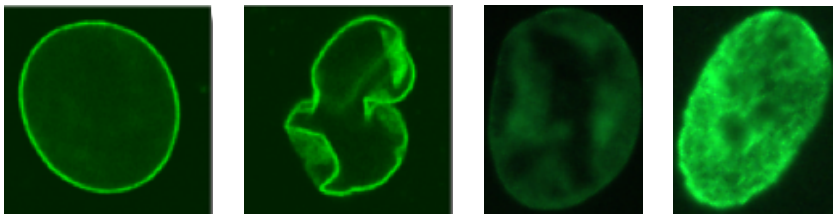
La reconnaissance de formes est une partie majeure de l'intelligence artificielle visant à automatiser le discernement de situations typiques au niveau de la perception. Elle est un enjeu primordial pour de très nombreuses applications : la reconnaissance des caractères manuscrits (numérisation des livres, lecture automatique des lettres postales et des chèques bancaires, etc.), la vidéo surveillance (reconnaissance faciale), l'imagerie médicale (échographie, scanner, imagerie par résonance magnétique, classement des chromosomes), la télédétection, etc. et dans ce qui nous intéresse ici : l'aide au diagnostic.

Au cœur de la reconnaissance de formes, il y a une première étape incontournable : la caractérisation de formes. En effet, afin de pouvoir reconnaître un objet ou un individu, il faut tout d'abord le décrire (Trier *et al.,* 1996) et donc définir ses caractéristiques (morphologiques, géométriques, texturelles, etc.), puis retrouver et identifier ces mêmes caractéristiques sur la source numérique à analyser. Pour cela il est souvent nécessaire d'étudier les objets selon deux critères : le premier est la forme avec des méthodes d'analyse globale ou de contour (Jain *et al.,* 2000, Mari *et al.,* 2004) et le deuxième est la texture (Tuceryan *et al.,* 1998).

Le but de nos travaux est d'élaborer un modèle de classement de noyaux de cellules de patients atteints par la maladie de la Progéria (également connue sous le nom de syndrome de Hutchinson-Gilford) dans les classes *noyau sain* ou *noyau pathologique*. Cette maladie orpheline (une centaine de cas dans le monde) de type laminopathie provoque un vieillissement accéléré du patient. En 2003, une avancée majeure de la recherche (Sandre-Giovannoli *et al.,* 2003, Eriksson *et al.,* 2003) a mis en exergue la cause de cette maladie : une mutation dans le gène *LMNA* sur le chromosome 1. La *lamine A* produite par le gène LMNA en conditions physiologiques est présente à la périphérie et à l'intérieur du noyau (figure 1a). Par assemblage avec d'autres lamines, cette protéine participe au maintien de la structure du noyau et de son enveloppe. Le gène muté responsable de la maladie produit une protéine anormale qui ne remplit plus son rôle de maintien structural. Ce problème de structure engendre des noyaux avec une forme (figure 1b) et une texture (répartition des lamines A) anormales (figure 1c et d). Ces anomalies génèrent des problèmes de division (mitose) responsables du vieillissement prématuré. Afin d'évaluer la progression de la maladie ou l'efficacité d'un protocole de soin, les experts analysent l'évolution du pourcentage de noyaux anormaux. D'après les répercutions de la présence du gène muté précédemment citées, il est nécessaire de diagnostiquer l'état d'un noyau en étudiant les deux aspects suivants : la forme (*normale* ou bien *boursouflée*) qui joue un rôle principal dans le diagnostic et l'homogénéité de la texture (*homogène* ou *non homogène*). Pour cela, nous disposons d'un ensemble de noyaux de cellules sanguines prélevées chez des patients, ainsi que d'une expertise de ces noyaux.

Pour observer les noyaux, les images sont acquises à l'aide d'un microscope à fluorescence (Leica DMR) couplé à une caméra Princeton-Roger et on utilise un marqueur de type FITC (*Fluoresceine Iso Thio Cyanate*) afin d'observer la répartition des lamines A/C. Une segmentation est ensuite réalisée en combinant un filtrage par

transformée de Fourier rapide (Chinga *et al.,* 2007) et un seuillage par maximisation de l'entropie (Pun, 1980).



(a) (b) (c) (d)

**Figure 1.** *Exemples de noyaux marqués avec le FITC : (a) un noyau sain, (b) un noyau boursouflé, (c) et (d) deux noyaux avec un texture non homogène*

Cet article montre les différentes étapes de la construction d'un modèle de classement des noyaux à partir de la caractérisation et du classement de leur forme et de l'homogénéité de leur texture. Dans un premier temps nous décrivons l'élaboration d'indices de formes dédiés à la caractérisation de la forme, puis la construction du sous-modèle de classement associé. Ensuite nous présentons une nouvelle méthode de caractérisation de la texture, basée sur la construction d'une matrice qui donne une représentation statistique de la texture. Les différentes méthodes de caractérisations employées sont systématiquement comparées en termes de performances à des techniques antérieures. Notre approche est confrontée à une méthode de classement utilisant un très vaste panel d'éléments de caractérisation de formes et de textures (Orlov *et al.,* 2008).

## 2. Le classement

Le but du classement est d'associer à chaque individu étudié une classe d'appartenance. Dans le problème qui nous concerne, l'objectif est de déterminer si un noyau de cellule est *sain* ou *pathologique*. Les méthodes de classement sont dites supervisées car elles font intervenir une expertise de référence. Dans notre cas, nous bénéficions de l'expertise des biologistes et généticiens qui ont déterminé les classes (*sain* et *pathologique*) et les sous-classes (forme *normale* et *boursouflée*, texture *homogène* et *non homogène*, etc.).

Un modèle pour le classement est souvent construit par apprentissage, à l'aide de données dont on connaît la classe *a priori*. Bien que s'appliquant à un problème spécifique, le modèle doit être capable de généraliser (au sens des données). Pour cela, les données sont séparées en deux groupes : un échantillon d'apprentissage et

un échantillon de validation. Le classifieur doit garder des performances comparables lors de l'apprentissage et de la validation.

Pour mettre en œuvre une méthode de classement, il faut au préalable construire un vecteur caractéristique de l'individu étudié. Le vecteur doit être pertinent au problème posé afin de permettre un bon apprentissage et une bonne prédiction. Le risque majeur lorsque l'on fournit trop de caractéristiques au classifieur est l'apprentissage *par cœur*. Plus la dimension du vecteur est grande, plus le modèle est adaptable et donc plus l'apprentissage est bon, mais plus la validation du modèle à l'aide d'individus non utilisés dans la phase de construction est mauvaise. Il faut alors systématiquement valider chaque modèle construit et obtenir le meilleur taux de classement sur l'échantillon de validation.

La principale méthode de classement choisie pour nos modèles est la *régression logistique* (Hosmer *et al.,* 1989). C'est un modèle linéaire particulièrement adapté pour les problèmes de classement à deux classes : $P = P(Y/\vec{x}) = \frac{e^{f(\vec{x})}}{1+e^{f(\vec{x})}}$ avec $\vec{x} = (x_1, \ldots x_n)$ le vecteur caractéristique de la donnée en entrée, $f(\vec{x}) = \sum_i \alpha_i x_i$ et $P(Y/\vec{x})$ la probabilité conditionnelle $P$ de la variable $\vec{x}$ d'appartenir à la classe $Y$. Pour estimer les coefficients $\alpha_i$ du modèle, on utilise le plus souvent la méthode du maximum de vraisemblance qui maximise la probabilité d'obtenir les valeurs observées sur l'échantillon d'apprentissage. Elle consiste à rechercher les paramètres qui optimisent la fonction de vraisemblance $L(\alpha, Y) = P^Y[1-P]^{1-Y}$. La régression logistique est préférée à l'analyse discriminante (Fisher, 1936) pour sa plus grande fiabilité, souplesse, le peu de restriction sur les variables et l'explicité de ses résultats. Toutefois nous utilisons trois autres méthodes de classement afin de comparer les performances et ainsi déterminer la plus adaptée :

– Les *k-plus proches voisins* (*k-nearest neighbor*) est une des plus anciennes, plus simples et plus intuitives méthodes de classement non linéaire (Fix *et al.,* 1951). Nous testons systématiquement différentes valeurs du paramètre $k$, qui peut être soit fixe, soit dépendant de la taille du vecteur caractéristique ($N_i + *$) ;

– Les *forêts aléatoires* (Breiman, 2001) (*random forests*) est une méthode de classement non linéaire basée sur l'utilisation d'arbres de décision (Morgan *et al.,* 1963) de type *Classification And Regression Trees* (CART) (Breiman *et al.,* 1984). C'est l'un des derniers aboutissements dans la recherche d'agrégation d'arbres de décision randomisés. Elle synthétise les approches développées dans (Breiman, 1996) et (Amit *et al.,* 1997) ;

– Les *réseaux de neurones* (McCulloch *et al.,* 1943). Ils sont très largement répandus grâce à leur puissance de modélisation (ils peuvent approcher n'importe quelle fonction suffisamment régulière), pour résoudre une grande variété de problèmes, face à des phénomènes complexes, des données difficiles à appréhender et ne suivant pas de lois probabilistes particulières. Nous utilisons un perceptron multicouche (Rosenblatt, 1958) avec une couche cachée dont le nombre de neurones dépend du nombre de neu-

rones de la couche d'entrée et de celle de sortie : $(N_{Input} + N_{Output})/\nu$. Nous testons systématiquement différentes valeurs de $\nu$ afin de trouver la meilleure architecture du réseau.

REMARQUE. — Concernant l'analyse des noyaux. Nous disposons de $2800$ noyaux expertisés, mais seulement $135$ possèdent une texture non homogène, soit environ $5\,\%$. Ce déséquilibre entre les deux classes « homogène » et « non homogène » ne permet pas de séparer les noyaux en deux échantillons équilibrés et représentatifs. Pour cela lors de l'étude de la texture, nous effectuons un sous-échantillonnage (*down-sizing* ou *under-sampling*) (Liu *et al.,* 2006), c'est-à-dire que nous réduisons la taille de la classe majoritaire en sélectionnant les $135$ noyaux les plus représentatifs (les parangons) parmi les noyaux à texture homogène. Puis nous les ajoutons aux $135$ noyaux à texture non homogène afin de construire l'échantillon de travail. Cette sélection s'effectue à l'aide d'une classification par k-moyennes (Hartigan *et al.,* 1979) qui est rendue robuste par le calcul préalable des formes fortes obtenues par la mise en œuvre successive de trois k-moyennes. Les parangons sont les noyaux les plus proches du barycentre de chaque forme forte. Nous aurions pu effectuer un sur-échantillonnage (*over-sampling*), en dupliquant de manière aléatoire ou dirigée les noyaux de la classe minoritaire (non homogène) jusqu'à équilibre des effectifs (Liu *et al.,* 2007). Mais il a été montré dans (Japkowicz, 2000, Liu *et al.,* 2006) que cette technique est *généralement* moins efficace. Il existe également d'autres techniques fondées sur la mesure d'entropie asymétrique (Marcellin *et al.,* 2006) ou sur l'utilisation d'un réseau de neurones auto-associateur (Japkowicz, 2000).

En raison de la faible taille de l'échantillon ($270$ individus), la validation est systématiquement effectuée à l'aide d'un protocole *leave one out* (Martens *et al.,* 1998) : de manière itérative, on construit un modèle à partir de tous les individus disponibles sauf un, ce dernier étant utilisé pour évaluer le modèle. La répétition de cette opération pour tous les individus de l'échantillon fournit le taux de classement.

## 3. Caractérisation et classement de la forme

L'analyse de la forme est l'élément le plus important dans le diagnostic des noyaux. En effet, l'étude de l'expertise révèle que $80\,\%$ des noyaux pathologiques ont une forme boursouflée. Il est donc nécessaire d'élaborer un sous-modèle de classement de la forme des noyaux. Pour cela, nous commençons par tester différentes méthodes de caractérisations de formes globales.

### 3.1. *Travaux antérieurs*

#### 3.1.1. *Les moments : Hu, Legendre et Zernike*

La notion de *moment* en mathématiques (notamment en calcul des probabilités) a pour origine la notion de moment en physique. Le moment $m_n(f)$ d'ordre $n = p + q$

d'une fonction $f$ définie sur un intervalle $\Upsilon$ (non réduit à un point) de $\mathbb{R}$ et ce même moment centré appliqué à une image $I$ contenant une forme $F$ sont :

$$m_n(f) = \int_{\Upsilon} x^n f(x) dx \qquad m_{pq}(I) = \sum_{(x,y)\in F} (x-\overline{x})^p (y-\overline{y})^q I(x,y)$$

Les moments apportent différentes informations statistiques sur la forme :

– ordre 0, surface de la forme : $m_{00}$.

– ordre 1, centre de gravité de la forme : $\overline{x} = \frac{m_{01}}{m_{00}}$ et $\overline{y} = \frac{m_{10}}{m_{00}}$

– etc.

Pour caractériser la forme des noyaux, nous utilisons trois familles de moments qui sont invariants par translation, rotation et homothétie :

– Les moments de *Hu* (Hu, 1962) sont 7 moments issus de produits et quotients des moments centrés normés d'ordre 3.

– Les moments de *Legendre* (Teague, 1980) sont des moments orthogonaux créés à l'aide du polynôme de Legendre qui forment une base orthogonale complète sur l'intervalle $[-1, 1]$.

– Les moments de *Zernike* (Zernike, 1934) sont des moments orthogonaux complexes définis sur un disque et basés sur une représentation polaire de la forme.

Afin de caractériser la forme, nous calculons les moments d'ordre $4$ de Legendre et Zernike, ainsi que les moments de Hu pour l'ensemble des noyaux. Nous testons les quatre méthodes de classement précédemment citées (cf. section 2) et réalisons une recherche exhaustive afin de déterminer le meilleur sous-ensemble de moments. Le meilleur résultat apporte un taux de classement de $81\,\%$ en utilisant les réseaux de neurones (avec $\nu = 3$) et une validation croisée, avec un sous-ensemble composé de 17 moments. Nous avons également testé les moments d'ordre 3 et 5, mais ces derniers fournissent des résultats de qualité inférieure.

### 3.1.2. *Histogrammes de projections*

La technique des histogrammes de projections (Zhou *et al.,* 2004, Lorigo *et al.,* 2006) est très répandue en reconnaissance des caractères. Elle renseigne sur l'épaisseur de la forme dans plusieurs directions. Chaque histogramme est calculé en comptant le nombre de pixels de la forme dans une direction $\delta$ : $HP(\delta) = \sum_F I_\delta(x,y)$. Cela revient à *projeter* les pixels de la forme dans une direction et à regarder les variations de la distribution marginale (figure 2). En deux dimensions, quatre directions de projections sont possibles : horizontale, verticale et deux diagonales. Cette méthode est insensible aux variations de type translation (la projection est translatée mais les valeurs inchangées). En revanche, elle n'est pas invariante par rotation, mais ce problème se résout en effectuant une rotation préliminaire suivant l'axe principal, ni par homothétie mais il y a un rapport constant entre les projections et il suffit de mettre les formes à la même échelle.

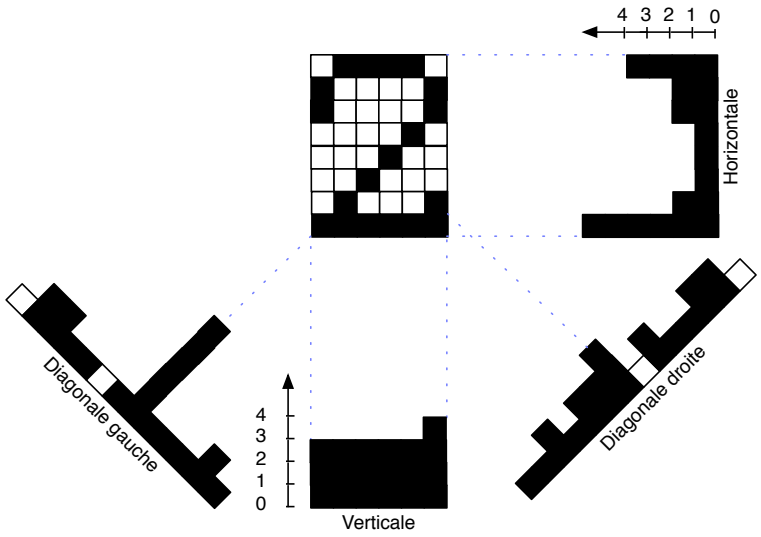



**Figure 2.** *Exemples d'histogrammes de projections horizontaux, verticaux et diagonaux pour le chiffre* 2

Nous employons les histogrammes de projections en utilisant les projections horizontales et verticales. Toutefois, il est nécessaire de *pré-traiter* les noyaux afin d'obtenir les invariances par rotation et homothétie. Pour cela, chaque noyau subit tout d'abord une rotation afin de confondre son axe principal avec l'axe des $X$, puis une transformation d'échelle afin que la boîte englobante soit de dimensions $N \times N$. Ainsi les histogrammes horizontaux et verticaux sont de taille $N$ et apportent $2N$ caractéristiques. Nous avons testé cette technique pour différentes valeurs de $N$ et ce, pour chaque méthode de classement. Le meilleur résultat est $83\,\%$ par régression logistique et validation croisée, pour un nombre de caractéristiques égal à 32 (dimensions $16 \times 16$). Bien que supérieur au résultat précédemment obtenu, ce pourcentage reste faible pour notre problème.

### 3.2. *Indices de formes*

Les indices de formes ont été présentés pour la première fois par Santalo (Santalo, 1976) dans un ouvrage relatif aux propriétés mathématiques des formes convexes. On trouve la définition et les propriétés des indices de formes dans (Coster *et al.,* 1985, Fillère, 1995).

**Définition 3.1 (Indice de formes)** *On appelle « indice de formes » tout paramètre, cœfficient ou combinaison de cœfficients permettant de donner des renseignements chiffrés sur la forme. De plus, les indices doivent avoir les propriétés suivantes :*

*1) Etre sans dimension ;*

*2) Etre invariants par homothétie ;*

*3) Etre invariants par rotation et translation.*

Le calcul d'un indice de formes est équivalent au calcul de la valeur d'une fonction à plusieurs variables, que l'on nomme *mesures* (figure 3).

**Définition 3.2 (Mesure)** *On appelle « mesure » d'une forme toute valeur ou ensemble de valeurs numériques « mesurées » sur la forme.*

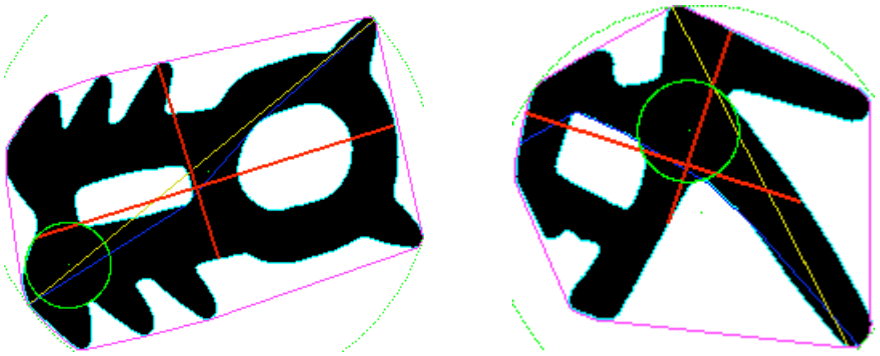


**Figure 3.** *Deux exemples de mesures : surface (noir), périmètre, axes principaux, enveloppe convexe, diamètre géodésique et diamètre euclidien, plus petite (resp. grande) boule circonscrite (resp. inscrite)*

L'extraction des mesures représente l'étape la plus importante dans le calcul d'un indice de formes, car de la valeur des mesures dépend la valeur de l'indice. Le temps de calcul et le comportement d'un indice dépendent de ceux des mesures qui le composent. Par exemple, le plus petit rayon est sensible au bruit de type *poivre et sel*, ce qui affecte de la même façon tout indice utilisant cette mesure. La liste complète des mesures et des indices utilisés dans notre travail est en annexe de cet article. Mais certaines mesures bien connues comme le diamètre de Ferêt ou le rayon de courbure maximum (resp. minimum) du contour ne sont pas utilisées car leur temps de calcul est trop important (complexité élevée).

Le principal avantage des indices de formes est leur grande souplesse. En effet, il est aisé de construire de nouveaux indices en fonction du problème que l'on souhaite traiter. Ces nouveaux indices spécifiques ont ainsi une grande capacité de description et permettent un meilleur classement. De plus, chaque indice apporte une valeur ou un ensemble de valeurs qui sont directement utilisables dans un classifieur.

### 3.2.1. *Indices de caractérisation d'une ellipse*

Les noyaux sains ont une forme allongée, régulière et quasi elliptique (figure 1a et figure 4a). C'est la forme résultante de l'aplatissement (lors de l'observation sur une

lame de microscope) d'une forme ellipsoïdale (forme naturelle d'un noyau) observée en 2D.

La mesure la plus simple à extraire d'une forme est l'aire, qui pour une ellipse se calcule de la façon suivante : $A = \pi ab$, avec $a$ le demi grand axe et $b$ le demi petit axe (figure 4). Or dans une ellipse on observe que :

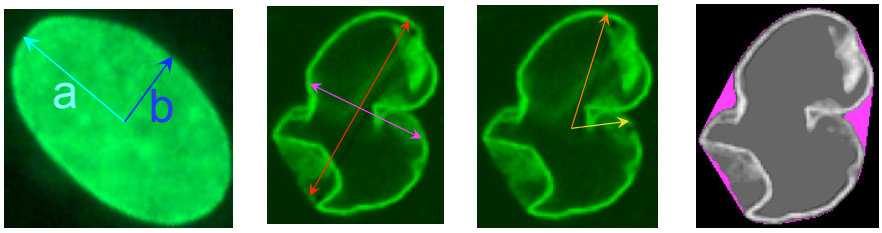
– Le grand axe est confondu avec l'axe principal et le diamètre ;

– Le demi grand axe est égal au plus grand rayon ;

– Le petit axe est porté par l'axe secondaire ;

– Le demi petit axe est égal au plus petit rayon et à l'épaisseur issue du diamètre.

Nous pouvons en déduire les égalités suivantes : $a = \frac{1}{2}L_{AP} = \frac{1}{2}D = R_{max}$ et $b = \frac{1}{2}L_{AP\perp} = E_D = R_{min}$ (figure 4).




(a) (b) (c) (d)

**Figure 4.** *(a) Illustration des demi axes sur un noyau avec une forme normale. (b) Axe principal et axe secondaire. (c) Plus grand rayon et plus petit rayon. (d) Illustration du calcul de la mesure $N_{Cce}$, on compte le nombre de composantes connexes d'écarts*

Ces inégalités permettent de construire trois nouveaux indices de caractérisation des ellipses, par :

**Les rayons**

$$\Psi_{EllipseR} = \pi \frac{R_{min}R_{max}}{A} \in [0,1]$$

**L'axe principal**

$$\Psi_{EllipseAP} = \frac{\pi}{4} \frac{L_{AP}L_{AP\perp}}{A} \in [0,1]$$

**Le diamètre**

$$\Psi_{EllipseD} = \frac{\pi}{2} \frac{E_D D}{A} \in [0,1]$$

L'indice d'ellipse par les rayons caractérise une ellipse en utilisant des mesures dépendantes du contour. En revanche, l'indice d'ellipse par l'axe principal utilise des mesures qui prennent en compte la totalité des points du noyau. Par construction, ces indices valent 1 pour des ellipses. Les intervalles d'appartenance sont calculés pour des formes convexes variant du segment au disque (Coster *et al.,* 1985).

3.2.2. *Indice de caractérisation de la convexité*

Nous venons de construire trois indices qui caractérisent les noyaux à forme normale. Il serait maintenant intéressant d'élaborer un indice pour caractériser les noyaux à forme anormale (*boursouflée*).

Le critère de décision principal dans le diagnostic de la forme des noyaux est la convexité. En effet, les noyaux boursouflés possèdent des zones de concavité en taille et en nombre différents. Pour compter ces zones de concavité, il est possible de calculer le *nombre de composantes connexes d'écart* $N_{Cce}$ issues de la soustraction de la forme à son enveloppe convexe $N_{Cce} = card(C_H(F)\backslash F)$ (figure 4d). Pour construire cet indice, nous utilisons la forme normée de la mesure :

$$\Psi_{N_{Cce}} = \frac{1}{1 + N_{Cce}} \in ]0, 1]$$

Cet indice vaut $1$ si la forme est convexe car aucune composante d'écart n'est trouvée et plus la forme possède des composantes d'écart, plus l'indice tend vers $0$.

Mais dans la pratique, on ne peut considérer comme composante connexe d'écart, des composantes dont la taille est de l'ordre du pixel et qui sont dues à des imprécisions de discrétisation. De plus, dans les éléments de diagnostic des noyaux, la taille et le nombre des composantes connexes doivent être pris en compte. Donc la mesure $N_{Cce}$ nécessite une calibration pour décider si une composante connexe d'écart doit être comptabilisée. Pour cela, nous avons réalisé une étude systématique du pourcentage de classement (obtenu en utilisant uniquement l'indice $\Psi_{N_{Cce}}$) en fonction de la taille et du nombre de composantes connexes d'écarts (figure 5). Cette étude consiste à calculer systématiquement le pourcentage de classement en faisant varier les seuils de taille et de nombre des composantes. C'est-à-dire que pour une taille $t$ et un nombre $n$, on ne comptabilise que les composantes dont la surface est plus grande que $t$, puis un noyau est considéré comme boursouflé si $N_{Cce}$ est supérieur ou égal à $n$. Il s'agit donc d'étudier une fonction discrète à deux variables qui produit une surface représentant le pourcentage de classement (figure 5).

Cette analyse met en exergue l'utilité de l'indice $\Psi_{N_{Cce}}$ dans le cas de noyaux non convexes ayant au minimum : soit une zone de concavité d'au moins trente-deux pixels, soit deux zones de concavité d'au moins douze pixels. On peut également remarquer une partie totalement plane sur la surface résultat. Elle correspond à des seuils en taille ou/et en nombre trop haut qui ont engendré le classement de tous les noyaux dans la classe « forme normale », ce qui montre que nous possédons près de $70\,\%$ de noyaux avec une forme normale.

Les deux seuils extraits de la surface résultat sont utilisés de manière équivalente dans $\Psi_{N_{Cce}}$ (sans pondération) et leur combinaison permet ainsi d'obtenir un taux de prédiction de plus de $90\,\%$ pour le sous-problème de forme avec ce seul indice.

Cet indice est utilisable dans tous les problèmes de caractérisation de la convexité. Il peut être employé directement sans calibration, mais celle-ci permet de l'adapter spécifiquement au problème traité. La calibration représente donc l'inconvénient de

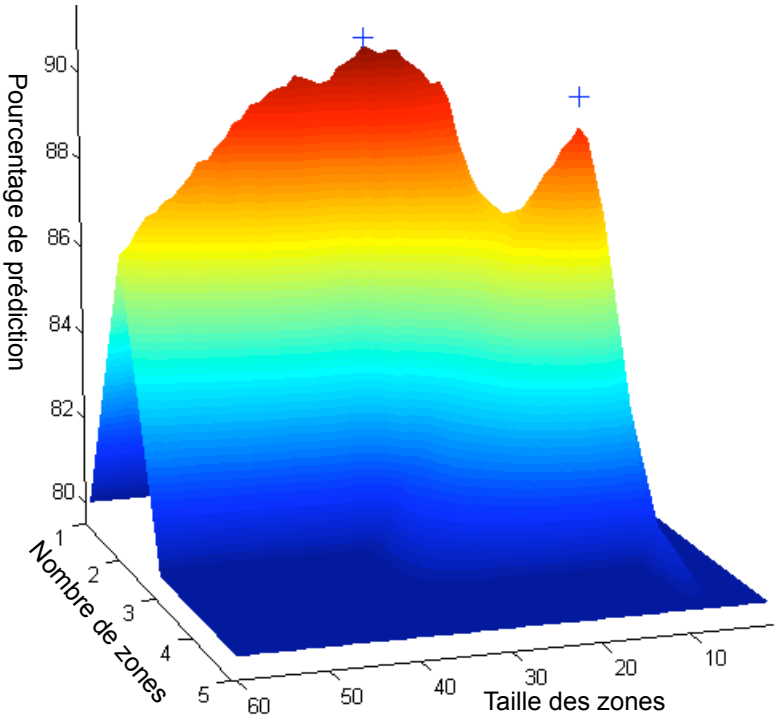



**Figure 5.** *Surface représentant le pourcentage de classement des noyaux en fonction du nombre et de la taille des composantes connexes d'écart*

cet indice en termes de complexité et d'étude, mais elle constitue également son principal avantage car elle améliore son efficacité.

### 3.3. *Sous-modèle de classement de la forme des noyaux*

Nous bénéficions de quatre nouveaux indices spécifiquement construits pour caractériser la forme des noyaux. A ces indices s'ajoutent treize indices sélectionnés dans la littérature scientifique, soit un total de dix-sept indices de formes. Pour chaque classifieur, nous avons effectué une recherche exhaustive (avec une validation croisée) afin de trouver le meilleur sous-ensemble d'indices (figure 6).

La régression logistique et le réseau de neurones apportent des résultats comparables. Ils permettent un pourcentage de classement de la forme des noyaux de $95,4\,\%$. La probabilité d'obtenir ce résultat de manière aléatoire est inférieure à $10^{-4}$ et l'intervalle de confiance à $5\,\%$ d'erreur est $[95,2\ldots95,6]$ (Gosh, 1979). Les performances des deux techniques étant comparables, nous construisons le

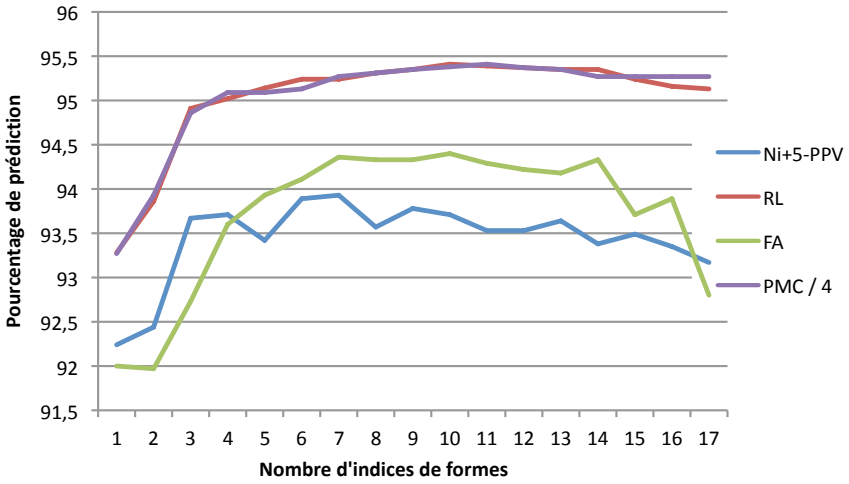



**Figure 6.** *Comparaison graphique des performances des méthodes de classement : plus proches voisins (PPV avec $N_i + 5$), régression logistique (RL), forêt aléatoire (FA) et perceptron multicouche (PMC avec $\nu = 4$). En abscisse le nombre d'indices de formes utilisés et en ordonnée le pourcentage de prédiction*

sous-modèle à l'aide de la régression logistique. Il est préférable d'utiliser un modèle plus simple (meilleure généralisation) et dont la probabilité associée à un individu est plus rapide à calculer.

Sans l'utilisation des quatre indices que nous avons créés, le meilleur résultat est un sous-ensemble composé de dix indices sur les treize disponibles. Il permet $93,6\,\%$ de prédiction avec un intervalle de confiance de $[93,4 \ldots 93,9]$. Les résultats sont $2\,\%$ plus faibles que ceux du modèle précédent et les deux intervalles de confiance sont disjoints. Ceci démontre l'efficacité des indices dédiés que nous avons élaborés.

Sur l'histogramme (figure 7) on remarque :

– La forte répartition des probabilités sur les extrémités de l'histogramme ;

– La quasi absence d'erreurs graves : des faux avec des probabilités extrêmes (inférieures à $0,2$ ou supérieures à $0,8$) ;

– La présence de seulement quelques cas ambigus : individus dont la probabilité est comprise entre $0,3$ et $0,7$ (proche de la valeur de décision $0,5$).

Ces trois informations montrent le peu d'ambiguïté dans le classement et par conséquent confirment le pouvoir de classement du sous-modèle ainsi que son efficacité. La liste suivante montre le meilleur sous-ensemble d'indices utilisés, classés en fonction du test du $\chi^2$ par ordre décroissant d'importance dans le sous-modèle :

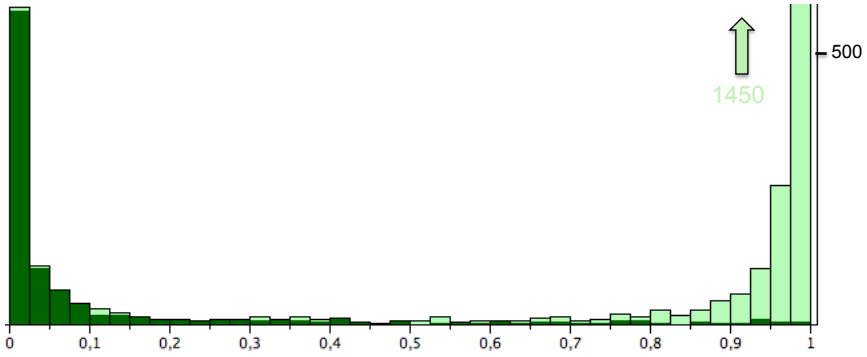



**Figure 7.** *Distribution des probabilités attribuées aux noyaux par le sous-modèle de classement de la forme. En gris clair (resp. gris foncé) les individus ayant une forme normale (resp. boursouflée)*

| $\chi^2$ | Indices | | |
|---|---|---|---|
| $0,476$ | $\Psi_{Ncce}$ | $0,018$ | $\Psi_{EllipseAP}$ |
| $0,345$ | Convexité surfacique | $0,017$ | Convexité périmétrique |
| $0,055$ | $\Psi_{EllipseD}$ | $0,012$ | Allongement rayons |
| $0,031$ | Symétrie de Besicovitch | $0,008$ | Déficit |
| $0,022$ | Allongement diamètre | $0,004$ | Circularité |

## 4. Caractérisation et classement de la texture

### 4.1. *Travaux antérieurs*

#### 4.1.1. *Indices de formes pour la caractérisation de textures*

Les indices de formes ont montré leur efficacité dans la construction du sous-modèle de classement de la forme. Cependant Chen et *al.* (1995) ont développé une approche qui permet de les utiliser pour caractériser des textures. Pour ce faire, ils considèrent une image $I$ comme la somme des résultats de tous les seuillages binaires :

$$I(x,y) = \sum_{\alpha=1}^{N} f_b(x,y,\alpha) \text{ avec } f_b(x,y,\alpha) = \begin{cases} 1 \; si \; I(x,y) \geqslant \alpha \\ 0 \; sinon \end{cases}$$

avec $N$ le nombre de niveaux de gris de l'image. Soit $E_\alpha$, l'ensemble des composantes connexes issues du seuillage de $I$ pour un seuil $\alpha \in [1,N]$. Pour chaque valeur de $\alpha$ et pour un indice de formes $\chi$, on peut calculer :

$$\chi_\alpha = \frac{\sum_{x \in E_\alpha} [S(x) * \chi(x)]}{\sum_{x \in E_\alpha} S(x)} \text{ avec } S(x) \text{ la surface de } x$$

Les quatre valeurs *max*, *moyenne*, *moyenne pondérée* et *SampleSD* calculables sur ces indices $\chi_\alpha$constituent des caractéristiques de la texture pour l'indice $\chi$.

Nous avons utilisé cette technique avec les dix indices employés dans le modèle de caractérisation de la forme des noyaux. Mais chaque indice produit quatre valeurs et nous avons été confronté au problème de l'apprentissage par cœur. De plus, aucun sous-ensemble d'indices ne s'est révélé performant pour répondre à notre problème. Le meilleur (composé de 28 indices) apporte un pourcentage de classement de $85\,\%$.

4.1.2. *Matrice de cooccurrences et caractéristiques Haralick*

La matrice de cooccurrences (ou matrice de dépendance spatiale) est une des approches les plus connues et les plus utilisées pour extraire des caractéristiques de textures. Elle effectue une analyse statistique de second ordre de la texture, par l'étude des relations spatiales des couples de pixels (Haralick *et al.,* 1973, Haralick, 1979).

La matrice de cooccurrences s'intéresse aux relations qui existent entre les niveaux de gris des pixels de la texture pour un déplacement (translation) $\overrightarrow{d}$ donné. Le résultat est une matrice carrée de taille $N \times N$, où $N$ est le nombre de niveaux de gris de la texture. Pour un déplacement $\overrightarrow{d} = (dx, dy)$, un élément $(x, y)$ de la matrice est défini par le nombre de pixels de la texture de niveau de gris $y$ situés à un déplacement $\overrightarrow{d}$ d'un pixel de niveau de gris $x$ (figure 8). Ce qui peut s'écrire formellement :

$$M_d(x,y) = card\left\{\left((r,s),(r+dx,s+dy)\right)/I\left(r,s\right)=x, I\left(r+dx,s+dy\right)=y\right\}$$

Texture

| | | | |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 0 | 0 | 2 | 2 |
| 0 | 0 | 2 | 2 |

$\Longrightarrow$

| $N$ | $\overrightarrow{d} = (0,1)$ | | | $\overrightarrow{d} = (1,1)$ | | |
|---|---|---|---|---|---|---|
| i \ j | 0 | 1 | 2 | 0 | 1 | 2 |
| 0 | 4 | 0 | 2 | 3 | 1 | 1 |
| 1 | 2 | 1 | 0 | 1 | 1 | 0 |
| 2 | 0 | 0 | 2 | 1 | 0 | 1 |

**Figure 8.** *Exemples de remplissage de la matrice de cooccurrences pour deux déplacements ($\overrightarrow{d} = (1,0)$ et $\overrightarrow{d} = (1,1)$) pour une texture à 3 niveaux de gris*

La matrice de cooccurrences met en évidence les relations qui existent entre les pixels à la fois par un aspect local (les niveaux de gris) et un aspect spatial (le déplacement). Cependant, toutes les caractéristiques sont extraites si on calcule un grand nombre de matrices : tous les déplacements combinés avec tous les niveaux de gris, ce qui génère une quantité importante d'informations. Pour un nombre de niveaux de gris $N$ préfixé, nous calculons quatre matrices pour les quatre déplacements $\overrightarrow{d_1} = (1,0)$, $\overrightarrow{d_2} = (1,1)$, $\overrightarrow{d_3} = (0,1)$ et $\overrightarrow{d_4} = (-1,1)$. Ensuite nous effectuons la moyenne des quatre matrices résultats (Wouwer *et al.,* 1999), ce qui permet de fusionner les informations et de s'abstraire de la direction. Toutefois, il est toujours nécessaire de faire varier le nombre de niveaux de gris et de tester des déplacements plus importants (éloignement $\varepsilon\overrightarrow{d_i}$, $\varepsilon \in \mathbb{N}^*$), ce qui augmente le temps de calcul et la

quantité d'information $M_{\varepsilon,N} = \frac{1}{4}\sum_{i=1}^{4} M_{\varepsilon\overrightarrow{d_i},N}$. A partir de cette matrice réduite, on extrait différents attributs appelés *indices de texture du second ordre* ou *caractéristiques Haralick* (Haralick *et al.,* 1973, Haralick, 1979) du nom de leur concepteur.

Nous réalisons une recherche exhaustive afin de déterminer la méthode de classement la plus efficace et le meilleur sous-ensemble d'indices de textures. De plus, nous effectuons les calculs pour $16$, $32$ et $64$ niveaux de gris et pour différents éloignements ($\varepsilon = 1 \ldots 5$). En raison du faible nombre d'individus dans l'échantillon de travail (cf. section 2), la validation est effectuée à l'aide du protocole *Leave One Out*.

Le meilleur sous-modèle de classement est construit par régression logistique (les réseaux de neurones apportent des résultats comparables). Il est calculé sur une matrice à $32$ niveaux de gris, pour un éloignement de $1$ et constitué de $8$ indices : variance, corrélation, moyenne des sommes, entropie des sommes, entropie, variance des différences, homogénéité et dissimilarité. Le sous-modèle obtient un pourcentage de classement de $89,8\,\%$ avec un intervalle de confiance de $[87,1 \ldots 92,5]$ et une probabilité inférieure à $10^{-4}$. Mais on remarque que le pourcentage de prédiction est nettement plus élevé que la borne supérieure de l'intervalle de confiance, ce qui implique que ce sous-modèle est sensible aux données : supprimer des données perturbe les performances. Malgré une importante répartition sur les extrémités (figure 9), on peut constater qu'il existe des cas ambigus ($40$ individus) et des erreurs graves ($8$ individus).

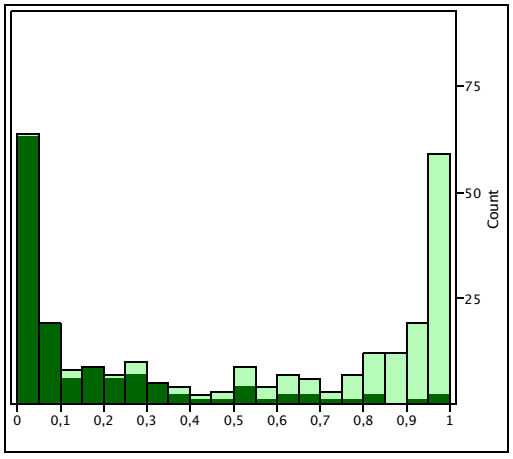



**Figure 9.** *Histogramme des distributions des probabilités attribuées par le sous-modèle de classement de l'homogénéité de la texture par caractéristiques Haralick. En gris clair (resp. gris foncé) les individus à texture homogène (resp. non homogène)*

4.1.3. *Gray level run length matrix*

La matrice de longueur de segments (*gray level run length matrix*) est une méthode statistique de caractérisation de la texture (Haralick *et al.,* 1973, Chu *et al.,* 1990). Cette méthode effectue le comptage du nombre de segments de pixels de même intensité dans une direction donnée et les résultats sont représentés dans une matrice. Pour cela, une direction ($0°$, $45°$, $90°$ ou $135°$) et un nombre de niveaux de gris $N$ sont préalablement fixés. La valeur contenue dans la case $(l, n)$ de la matrice est égale au nombre de segments de longueur $l$ et de niveaux de gris $n$. Donc le nombre de colonnes de la matrice est variable car il dépend de la longueur du plus long segment. De par sa conception, le calcul est symétrique, il est par conséquent inutile de la calculer dans les quatre directions complémentaires ($180°$, $225°$, $270°$ ou $315°$, on considère ici huit directions possibles entre le pixel étudié et ses voisins). La figure 10 montre un exemple de remplissage de la matrice.

Texture

| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 1 | 3 | 4 | 4 |
| 3 | 2 | 2 | 2 |
| 4 | 1 | 4 | 1 |

$\Longrightarrow$

| $N$ | Run length (j) | | | |
|---|---|---|---|---|
| i | 1 | 2 | 3 | 4 |
| 1 | 4 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 3 | 0 | 0 | 0 |
| 4 | 3 | 1 | 0 | 0 |

**Figure 10.** *Exemple de remplissage de la matrice « run length » pour une texture* $4 \times 4$ *à quatre niveaux de gris et dans la direction* $0°$

Onze indices sont calculés (Xu *et al.,* 2004) afin de construire le vecteur caractéristique de la texture. Pour construire notre modèle, nous calculons ces caractéristiques pour un niveau de gris fixé et dans les quatre directions. Puis pour chaque indice, on calcule la moyenne de ses valeurs dans les quatre directions. Une recherche exhaustive a montré que le meilleur sous-modèle est obtenu pour un ensemble de 7 indices avec 32 niveaux de gris. Le taux de classement obtenu est de $84,81\,\%$ par régression logistique ou réseaux de neurones, ce qui est un résultat moins performant que celui obtenu avec la matrice de cooccurrences et les caractéristiques Haralick ($90\,\%$).

## 4.2. *Gray level size zone matrix*

Une texture homogène possède de grandes zones de même intensité et non pas des segments dans une direction donnée. Pour tenir compte de cette remarque, nous proposons une approche qui comptabilise toutes les tailles des zones de pixels de même niveau d'intensité dans une matrice. Cette dernière est construite sur le principe de la *Run Length Matrix* : la valeur de la case $(s, n)$ de la matrice contient le nombre de zones de taille $s$ et de niveau de gris $n$. La figure 11 montre un exemple de remplissage de cette matrice, baptisée *size zone matrix*.

La matrice produite possède un nombre de lignes fixe égal au nombre de niveaux de gris et un nombre de colonnes variable qui dépend de la taille de la plus grande

Texture

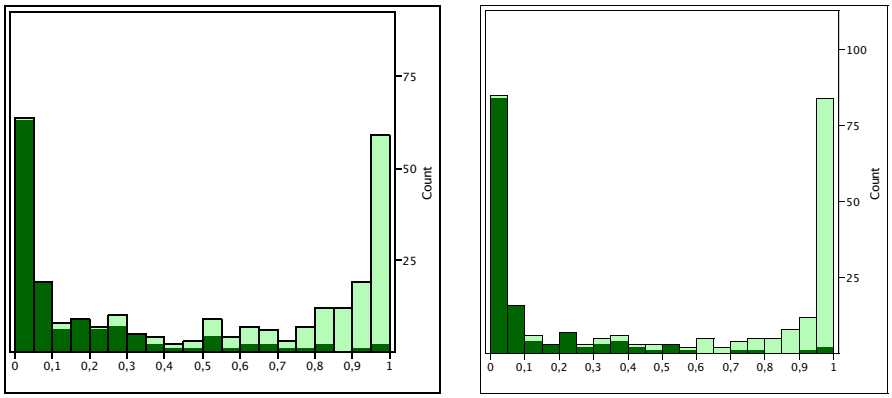
| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 1 | 3 | 4 | 4 |
| 3 | 2 | 2 | 2 |
| 4 | 1 | 4 | 1 |

$\Longrightarrow$

| $N$ | Size zone (j) | | | |
|---|---|---|---|---|
| i | 1 | 2 | 3 | 4 |
| 1 | 2 | 1 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 | 0 |
| 4 | 2 | 0 | 1 | 0 |

**Figure 11.** *Exemple de remplissage de la matrice « size zone » en* $8$*-connexités pour une texture* $4 \times 4$ *à quatre niveaux de gris*

zone. Plus la texture est homogène, plus la matrice est *large* et *creuse*. Cette matrice possède l'avantage de ne pas nécessiter de calculs dans plusieurs directions, qui sont remplacés par un étiquetage des différentes zones. En revanche, le fait de spécifier le nombre de niveaux de gris rend le remplissage robuste au bruit. En $32$ niveaux de gris, nous calculons ensuite les mêmes onze indices que pour la matrice *run length*. Le pourcentage de classement pour les onze indices est de $91,11\,\%$ par régression logistique, l'intervalle de confiance est $[89,1\ldots93,1]$ et la probabilité est inférieure à $10^{-4}$. On peut remarquer sur la figure 12b la forte concentration des probabilités vers les extrémités de l'histogramme, la présence de seulement $29$ cas ambigus et $6$ erreurs graves. Cette répartition illustre la puissance de classement du sous-modèle. Ces résultats montrent que cette méthode de caractérisation de la texture des noyaux est la meilleure parmi toutes celles testées jusqu'alors.




(a) (b)

**Figure 12.** *Comparaison des distributions des probabilités de classement données par les sous-modèles basés sur les caractéristiques Haralick (a) et la Glszm (b). En gris foncé (resp. gris clair) les noyaux à texture non homogène (resp. homogène). Plus la probabilité est proche de* $1$ *(resp.* $0$*), plus la texture est homogène (resp. non homogène)*

Toutefois en étudiant les résultats, il apparaît une similitude entre certains faux positifs : les noyaux possèdent de grandes zones homogènes, mais avec de fortes variations d'intensité entres les zones (figure 13), ce qui en font des noyaux à texture non homogène.

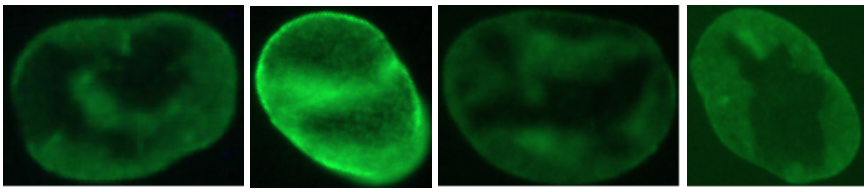


**Figure 13.** *Exemples de faux positifs : des noyaux à texture non homogène classés parmi les noyaux à texture homogène. Des noyaux avec des grandes zones homogènes, mais avec des variations d'intensité importantes*

Pour caractériser ce type de noyaux, nous introduisons deux nouveaux indices basés sur l'écart type pondéré par les niveaux de gris ou les tailles des zones :

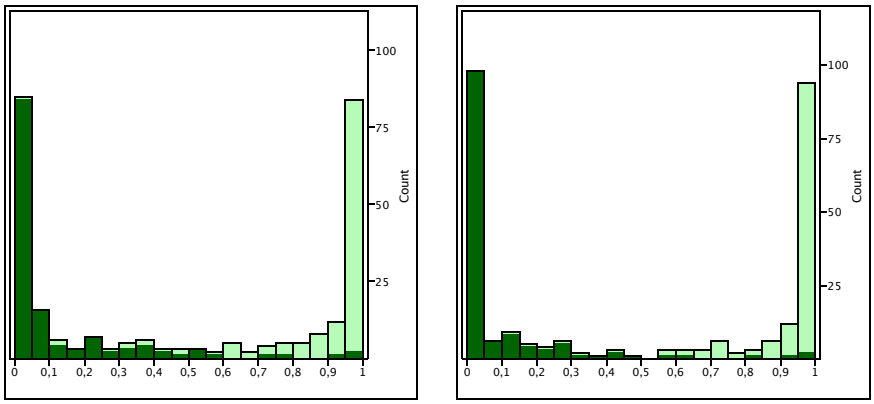
$$\Psi_N = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(n*M(n,s)-\mu_N\right)^2}\ ,\ \mu_N = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} n*M(n,s)$$

$$\Psi_S = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(s*M(n,s)-\mu_S\right)^2}\ ,\ \mu_S = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} s*M(n,s)$$

avec $N$ et $S$ les dimensions de la matrice et $M(n,s)$ l'élément de coordonnées $(n,s)$ de la matrice. Plus la texture contient de grandes zones avec des écarts d'intensité importants, plus la valeur de l'indice $\Psi_N$ est grande. Une texture homogène engendre une valeur faible de cette indice. Il en est de même pour l'indice $\Psi_S$ avec les écarts entre les tailles des zones.

Grâce à l'utilisation de ces deux indices dans le modèle composé des douze indices les plus pertinents (tous sauf *Low Run High Gray*), le taux de classement est désormais de $94,07\,\%$ par régression logistique (figure 15). Le réseau de neurones avec $\nu = 2$ obtient des résultats comparables. L'intervalle de confiance est $[92,1\ldots96,1]$ et la probabilité du sous-modèle est inférieure à $10^{-4}$. L'utilisation de nos deux indices permet d'améliorer la prédiction du sous-modèle de près de $3\,\%$.

Il existe une intersection entre l'intervalle de confiance de ce dernier sous-modèle et celui du précédent n'utilisant pas nos deux indices. Une étude de rang en fonction de la médiane révèle que la probabilité du sous-modèle précédent d'apporter des résultats équivalents au nouveau sous-modèle est de $0,087$. De même, une analyse de la variance (test de Wilcoxon (Tufféry, 2007, Wonnacott *et al.,* 1998)) des taux de classement utilisés pour le calcul des intervalles de confiance de chaque sous-modèle, révèle que la probabilité de ces deux distributions d'être issues d'une même loi est




(a) (b)

**Figure 14.** *Comparaison des distributions des probabilités de classement données par les sous-modèles basés sur la Glszm : (a) indices classiques et (b) indices classiques et nos deux indices. En gris foncé (resp. gris clair) les noyaux à texture non homogène (resp. homogène). Plus la probabilité est proche de* $1$ *(resp.* $0$*), plus la texture est homogène (resp. non homogène)*

inférieure à $10^{-4}$. Ces trois informations montrent la nécessité d'utiliser nos deux indices.

Le gain apporté au sous-modèle par nos indices peut être observé sur la figure 14 *b*. La forte répartition sur les extrémités de l'histogramme montre l'efficacité de classement et la pertinence dans le choix des indices. On peut remarquer une nette amélioration de l'augmentation des distributions des probabilités vers les valeurs extrêmes et la faible présence de cas ambigus (seulement $16$ individus).

Afin de comparer et valider la pertinence de notre travail, nous faisons une comparaison avec une méthode récente et généraliste de classification (Orlov *et al.,* 2008) : la méthode extrait $2\,700$ caractéristiques de formes et de textures. Puis une sélection automatique du meilleur sous-ensemble de caractéristiques est réalisée, suivie d'une pondération des caractéristiques sélectionnées. Finalement un classement par k-plus proches voisins est opéré.

Nous avons testé cette méthode sur ce sous-problème de classement de la texture, car elle utilise un très large éventail de caractéristiques de textures, ce qui nous permet de situer l'efficacité de notre contribution. La meilleure configuration de cette méthode (utilisation de $10\,\%$ des caractéristiques les plus pertinentes) apporte un pourcentage de classement de $90\,\%$. Ce résultat est comparable à celui obtenu à l'aide des caractéristiques Haralick qui sont justement utilisées dans cette méthode. Ce dernier résultat montre donc la pertinence et l'efficacité de notre contribution (les Glszm) dans la résolution du sous-problème posé.

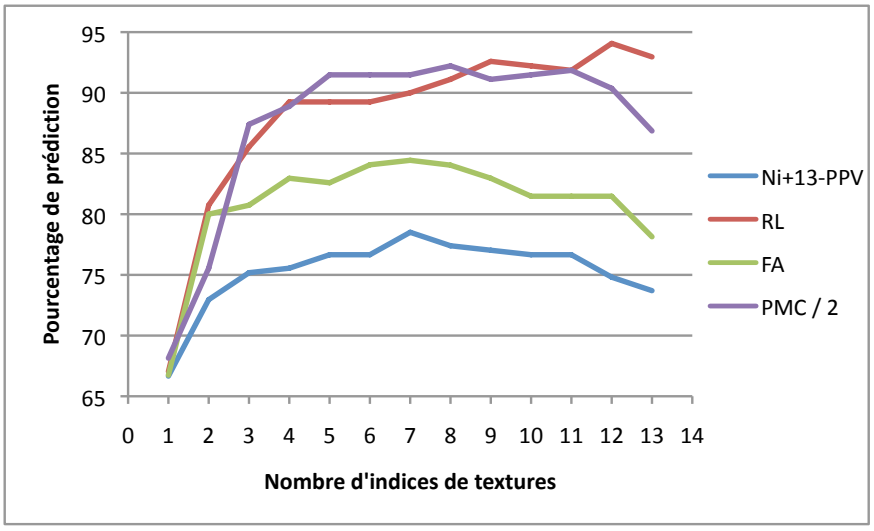



**Figure 15.** *Comparaison des performances des classifieurs appliqués au sous-problème de la texture : plus proches voisins (PPV avec $N_i+13$), régression logistique (RL), forêt aléatoire (FA) et perceptron multicouche (PMC avec $\nu = 2$). En abscisse le nombre d'indices de textures utilisés et en ordonnée le pourcentage de prédiction*

## 5. Conclusion

Nous venons de détailler la construction de deux sous-modèles de classement de la forme (cf. section 3.3) et de la texture (cf. section 4.2) des noyaux de cellules. Dans un premier temps, nous avons élaboré un sous-modèle de classement de la forme basé sur l'utilisation d'indices de formes dédiés qui se sont révélés particulièrement discriminants et donc efficaces. Par la suite, notre contribution a également apporté une nouvelle méthode de caractérisation statistique de l'homogénéité de la texture. Celle-ci est basée sur la construction de matrices représentatives de la texture dont l'efficacité a été comparée à une méthode utilisant un très grand nombre d'éléments de caractérisation de la texture.

Pour construire le modèle final de classement, nous utilisons la régression logistique (le réseau de neurones apportant des résultats comparables) afin de classer les noyaux en fonction des probabilités de classement de la forme et de la texture attribuées par les deux sous-modèles. Nous obtenons un pourcentage de prédiction de $87,8\,\%$. Ce résultat est inférieur à ceux des sous-modèles. Cela s'explique par le fait que la majorité des noyaux à texture non homogène (environ une centaine sur les $135$) ont une forme boursouflée. Le diagnostic de la texture est fortement corrélé à celui de la forme. De plus, seulement $94\,\%$ des noyaux sont diagnostiquables en fonction de leur forme et/ou de leur texture. Il persiste $6\,\%$ de noyaux dont l'état ne peut être diagnostiqué qu'en fonction d'informations survenant de manière occasionnelle. De

plus, nous avons appliqué la méthode décrite dans (Orlov *et al.,* 2008) au problème final de classement des noyaux. Celle-ci ne permet de classer que $71\,\%$ des individus, comparé au pourcentage que nous obtenons : $88\,\%$ sur les $94\,\%$ possibles.

Dans cet article, nous avons décrit les différentes étapes nécessaires à la réalisation d'un modèle de classement des noyaux de cellules prélevées chez des patients atteints par le syndrome de Hutchinson-Gilford. Grâce à notre modèle, il est désormais possible de classer de manière automatique, fiable et rapide les noyaux de cellules. Cette automatisation et ce gain de temps (par rapport au temps nécessaire au classement manuel des noyaux par les experts) constituaient l'objectif de notre travail qui a donc été atteint. De plus, notre travail a montré la faisabilité du classement automatique de noyaux de cellules par l'étude de la répartition des lamines A et C.

Par ailleurs, des défauts similaires des lamines A et C ont été décelés chez des patients atteints par certaines formes de cancers ou des patients séropositifs sous trithérapie. Il a été montré qu'après une certaine période sous trithérapie, les patients présentent un vieillissement physique, physiologique et intellectuel d'une dizaine d'années. Notre travail va donc être appliqué à l'étude de ces pathologies, afin d'étudier les impacts des trithérapies et de certaines chimiothérapies sur les noyaux cellulaires.

## 6. Bibliographie


Amit Y., Geman D., « Shape quantization and recognition with randomized trees », *Neural computation*, vol. 9, n° 7, p. 1545-1588, 1997.

Breiman L., « Bagging predictors », *Machine Learning*, vol. 24, n° 2, p. 123-140, 1996.

Breiman L., « Random forests », *Machine Learning*, vol. 45, n° 1, p. 5-32, 2001.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J., *Classification And Regression Trees*, CRC Press, 1984.

Chinga G., Johnsen P. O., Dougherty R., Berli E. L., Walter J., « Quantification of the 3D microstructure of SC surfaces », *Journal of Microscopy*, vol. 227, p. 254-265, September, 2007.

Chu A., Sehgal C. M., Greenleaf J. F., « Use of gray value distribution of run lengths for texture analysis », *Pattern Recognition Letters*, vol. 11, n° 6, p. 415-419, 1990.

Coster M., Chermant J.-L., *Précis d'analyse d'images*, Editions du CNRS, 1985.

Eriksson M., Brown T. W., Gordon L., Glynn M. W., Singer J., Scott L., Erdos M. R., Robbins C. M., Moses T. Y., Berglund P., Dutra A., Pak E., Durkin S., Csoka A. B., Boehnke M., Glover T. W., Collins F. S., « Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome », *Nature*, vol. 423, p. 6937, April, 2003.

Fillère I., Outils mathématiques pour la reconnaissance de formes, PhD thesis, Université de St Etienne, September, 1995.

Fisher R. A., « The use of multiple measurements in taxonomic problems », *Annals of Eugenics*, vol. 7, p. 179-188, 1936.

Fix E., Hodges J., Discriminatory Analysis : Nonparametric Discrimination : Consistency Properties, Technical Report n° 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Gosh B. K., « A comparison of some approximate confidence intervals for the binomial parameter », *Journal of the American Statistical Association*, vol. 74, p. 894-900, 1979.

Haralick R. M., « Statistical and structural approaches to texture », *Proceedings of the IEEE*, vol. 67, p. 786-804, May, 1979.

Haralick R. M., Shanmugam K., Dinstein I., « Textural features for image classification », *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, n° 6, p. 610-621, 1973.

Hartigan J. A., Wong M. A., « A K-Means Clustering Algorithm », *Applied Statistics*, vol. 28, n° 1, p. 100-108, 1979.

Hosmer D., Lemeshow S., *Applied Logistic Regression*, John Wiley & Sons, Toronto, 1989.

Hu M.-K., « Visual pattern recognition by moment invariant », *IEEE Transactions on Information Theory*, vol. 8, n° 2, p. 179-187, February, 1962.

Jain A. K., Duin R. P. W., Mao J., « Statistical Pattern Recognition : A Review », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 1, p. 4-37, 2000.

Japkowicz N., « Learning from Imbalanced Data Sets : A Comparison of Various Strategies », *AAAI Workshop on Learning from Imbalanced Data Sets*, AAAI Press, p. 10-15, 2000.

Liu A., Ghosh J., Martin C., « Generative oversampling for mining imbalanced datasets », *in* , R. Stahlbock, , S. F. Crone, , S. Lessmann (eds), *International Conference on Data Mining*, CSREA Press, p. 66-72, 2007.

Liu X.-Y., Wu J., Zhou Z.-H., « Exploratory Under-Sampling for Class-Imbalance Learning », *IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC, USA, p. 965-969, December, 2006.

Lorigo L. M., Govindaraju V., « Off-Line Arabic Handwriting Recognition : A Survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, p. 712-724, May, 2006.

Marcellin S., Zighed D.-A., Ritschard G., « An asymmetric entropy measure for decision trees », *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, p. 1292-1299, 2006.

Mari J.-L., Sequeira J., « Closed Free-Form Surface Geometrical Modeling - A New Approach with Global and Local Characterization », *International Journal of Image and Graphics (IJIG)*, vol. 4, n° 2, p. 241-262, April, 2004.

Martens H., Dardenne P., « Validation and verification of regression in small data sets », *Chemometrics and intelligent laboratory systems*, vol. 44, n° 1-2, p. 99-121, 1998.

McCulloch W. S., Pitts W., « A logical calculus of the ideas immanent in nervous activity », *Bulletin of Mathematical Biophysics*, vol. 5, p. 115-133, 1943.

Morgan J. N., Sonquist J. A., « Problems in the Analysis of Survey Data, and a Proposal », *Journal of the American Statistical Association*, vol. 58, p. 415-435, 1963.

Orlov N., Shamir L., Macura T., Johnston J., Eckley D., Goldberg I., « Multi-purpose image classification using compound image transforms », *Pattern Recognition Letters*, vol. 29, n° 11, p. 1684-1693, October, 2008.

Pun T., « A new method for grey-level picture thresholding using the entropy of the histogram », *Signal Processing*, vol. 2, p. 223-237, 1980.

Rosenblatt F., « The Perceptron : probabilistic model for information storage and organization in the brain », *Psychological Review*, vol. 65, p. 386-408, 1958.

Sandre-Giovannoli A. D., Bernard R., Cau P., Navarro C., Amiel J., Boccaccio I., Lyonnet S., Stewart C. L., Munnich A., Merrer M. L., Levy N., « Lamin A truncation in progeria », *Science*, vol. 300, n° 5628, p. 2055, April, 2003.

Santalo L., *Integral Geometry and Geometric Probability*, Addison Wesley, 1976.

Teague M. R., « Image analysis via the general theory of moments », *Journal of the Optical Society of America (1917-1983)*, vol. 70, n° 8, p. 920-930, August, 1980.

Trier Ø. D., Jain A. K., Taxt T., « Feature extraction methods for character recognition - A survey », *IEEE Transactions on Pattern Recognition Letters*, vol. 29, p. 641-662, April, 1996.

Tuceryan M., Jain A. K., *Texture analysis*, 2nd edition edn, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1998.

Tufféry S., *Data Mining et statistiques décisionnelles*, 2nd edn, Edition Technip, June, 2007.

Wonnacott T. H., Wonnacott R. J., *Statistique : Economie - Gestion - Sciences - Médecine (avec exercices d'application)*, 4th edn, Economia, March, 1998.

Wouwer G. V. D., Scheunders P., Dyck D. V., « Statistical texture characterization from discrete wavelet representations », *IEEE Transactions on Image Processing*, vol. 8, n° 4, p. 592-598, April, 1999.

Xu D., Kurani A., Furst J., Raicu D., « Run-Length Encoding For Volumetric Texture », *International Conference on Visualization, Imaging and Image Processing (VIIP)*, p. 452-458, 2004.

Zernike F., « Diffraction theory of the cut procedure and its improved form, the phase contrast method », *Physica*, vol. 1, p. 689-704, 1934.

Zhou Z.-H., Geng X., « Projection Functions for Eye Detection », *Pattern recognition*, vol. 37, n° 5, p. 1049-1056, May, 2004.

## Annexe : liste des mesures et indices de formes

| | |
|---|---|
| **Allongement par le diamètre** | $E_D/D \in [0,1]$ |
| **Allongement par les rayons** | $\rho_i/\rho_e \in [0,1]$ |
| **Allongement géodésique** | $4A/(\pi D_G^2) \in [0, \frac{\pi}{4}]$ |
| **Circularité** | $R_{min}/R_{max} \in [0,1]$ |
| **Convexité périmétrique** | $P(C_H)/P \in ]0,1]$ |
| **Convexité surfacique** | $A/A(C_H) \in ]0,1]$ |
| **Déficit** | $\pi(\rho_e - \rho_i)^2/P^2 \in [0, \frac{\pi^2}{16}]$ |
| **Déficit iso-périmétrique** | $4\pi A/P^2 \in [0,1]$ |
| **Ecart au disque inscrit** | $\pi\rho_i^2/A \in [0,1]$ |

| | |
|---|---|
| **Etalement de Morton** | $4A/(\pi L_{AP}^2) \in [0,1]$ |
| **Irrégularité** | $(A + \sqrt{\pi}\max_{p\in F} d(p,B))/\sqrt{A}$ |
| **Symétrique de Bezicovitch** | $\max_{p\in F}\ A\left(F \bigcap Symétrique_p\left(F\right)\right)$ |
| **Variance circulaire** | $\frac{1}{\lvert Contour(F)\rvert \mu_R^2} \sum_{p\in Contour(F)} \left(\lVert p-B\rVert - \mu_r\right)^2$ |

avec :

$A$ l'aire.

$B$ le barycentre. Bien que le barycentre ne soit pas une mesure à proprement parler, ce dernier est souvent utilisé dans le calcul des mesures.

$C_H$ l'enveloppe convexe.

$D$ le diamètre.

$D_G$ le diamètre géodésique.

$E_D$ l'épaisseur issue du diamètre.

$L_{AP}$ la longueur de l'axe principal. On note également $L_{AP\perp}$ la longueur de l'axe secondaire orthogonal à l'axe principal.

$P$ le périmètre.

$\rho_e$ le plus petit disque circonscrit à la forme.

$\rho_i$ le plus grand disque inscrit dans la forme.

$R_{min}$ le plus petit rayon.

$R_{max}$ le plus grand rayon.

$\mu_r$ le rayon moyen.