

Analyse fonctionnelle de données dans les images

Kevin Wang¹ et Cédric Beaulac^{2*}

¹ Département de mathématiques et statistique, UdeM, Canada

² Département de mathématiques, UQAM, Montréal, Canada

* Auteur correspondant, email: beaulac.cedric@uqam.ca

Résumé

Depuis des années, les chercheurs doivent analyser des images et cela est un défi de taille, car la technologie de collectes d'images est en évolution constante. De plus, ce ne sont pas toutes les personnes qui possèdent les technologies nécessaires pour analyser des données de cette sorte. En conséquence, une approche proposée est d'analyser les images en tenant compte du contour et de la forme des objets dans celles-ci en utilisant l'analyse de données fonctionnelles (ADF). Pour tester l'approche proposée, nous avons utilisé le jeu de données HAM10000 qui porte sur l'étude de grains de beauté où l'objectif est de différencier les grains de beauté bénins de ceux de type cancérigènes. Avec les résultats obtenus, les méthodes proposées d'analyse de contours s'avèrent efficaces.

Table des matières

1	Introduction	3
2	Motivation biologique et mathématique	3
3	Description des données analysées	4
3.1	Méthodes de prétraitement	4
3.1.1	Retrait des poils	5
3.1.2	Sélection de canaux pour la segmentation	5
4	Méthodologie	6
4.1	Fondements mathématiques de l'ADF	6
4.2	Représentation fonctionnelle des contours	7
4.3	Modèle de déformation des contours	7
4.4	Méthodes algorithmiques	8
4.4.1	Extraction des contours par Marching Squares	8
4.4.2	Lissage des données	8
4.4.3	Alignement des contours par ICF	9
4.4.4	Analyse fonctionnelle en composantes principales (FPCA)	9
4.4.5	Classification par régression logistique pondérée	10
4.5	Stratégies de validation	11
4.6	Métriques de validation	11
5	Résultats	12
5.1	Analyse des erreurs	12
6	Discussion	13
7	Conclusion	14
7.1	Synthèse des résultats	14
7.2	Limites et défis	15
7.3	Perspectives futures	15

1 Introduction

L'analyse de formes dans les images (pixels) est un enjeu en vision par ordinateur et en analyse de données, avec des applications dans plusieurs domaines comme l'imagerie médicale. En effet, l'analyse des grains de beauté constitue un exemple pertinent pour ce type de problèmes, où les images permettent de détecter la présence de mélanomes, c'est-à-dire des cancers de la peau. À l'aide des images récoltées, la capacité à extraire et à décrire des formes géométriques permet de mieux de comprendre les phénomènes complexes et de faciliter certaines analyses.

Cependant, les données provenant des images contiennent souvent une grande variabilité, du bruit et des structures complexes qui rendent leur analyse difficile. En effet, les grains de beauté peuvent être de forme circulaire, mais ils peuvent également présenter d'autres formes plus complexes. Les méthodes utilisées par les scientifiques manquent de robustesse ou de précision quand elles sont appliquées à des formes hétérogènes, comme celles des grains de beauté [15].

Dans ce contexte, l'analyse des données fonctionnelles (ADF) semble très prometteuse, car elle permet d'améliorer la stabilité et la performance des résultats tout en réduisant la sensibilité aux variabilités [16]. Les grains de beauté ont la particularité de prendre des formes propres à chaque individu, ce qui rend leur classification médicale plus difficile.

Ce projet vise principalement à montrer que l'utilisation de l'ADF permet de différencier les grains de beauté bénins de ceux de type cancéreux. Autrement dit, il exploite cette branche des mathématiques statistiques à l'aide de notions d'apprentissage automatique, notamment l'analyse en composantes principales, afin de classer les grains de beauté. De plus, nous souhaitons étudier la sensibilité des analyses quant au choix de la forme de référence.

Ce rapport propose d'abord les objectifs de recherche ainsi que les données utilisées. Ensuite, la méthodologie, les résultats et leur discussion seront présentés. Enfin, une conclusion portant sur les limites et les perspectives de recherche sera proposée.

2 Motivation biologique et mathématique

Dans le domaine médical, les médecins utilisent différentes méthodes afin de visualiser et d'analyser les grains de beauté pour déterminer leur nature. Cependant, les grains de beauté présentent une grande diversité morphologique, ce qui rend leur caractérisation difficile. En effet, leur géométrie biologique peut être complexe, puisque leur forme globale peut être circulaire ou ovale, présenter des asymétries selon plusieurs axes et être influencée par différents facteurs, tels que le bruit provenant des images ou des détails microscopiques associés aux mélanocytes.

La caractérisation des grains de beauté repose notamment sur l'analyse de plusieurs

critères morphologiques. Les lésions bénignes peuvent présenter une structure relativement régulière, avec notamment une organisation bipolaire ou tripolaire et des motifs pigmentaires particuliers, comme des réseaux ou des prolongements rappelant des extensions. À l'inverse, lors de la transformation en mélanome, certaines modifications morphologiques peuvent apparaître, notamment une augmentation de l'hétérogénéité de la forme, une asymétrie accrue et des changements dans l'organisation cellulaire.

Au cours de la progression d'un mélanome, les cellules peuvent présenter des modifications phénotypiques associées à une perte de différenciation normale. La tyrosinase, une enzyme impliquée dans la synthèse de la mélanine, peut notamment être affectée dans certaines populations cellulaires. À un stade avancé, les lésions peuvent présenter des formes plus allongées ou fusiformes, avec une augmentation de la densité cellulaire et une hétérogénéité morphologique persistante [7].

Par ailleurs, les mathématiciens utilisent la géométrie fractale afin de caractériser des structures complexes présentant des irrégularités, des asymétries et des propriétés géométriques particulières. Puisque les mélanomes peuvent présenter des contours irréguliers, l'étude d'approches basées sur la géométrie fractale représente également une piste intéressante pour analyser leur morphologie [13].

3 Description des données analysées

Les données utilisées dans cette analyse proviennent de la base de données HAM10000, qui a été utilisée dans le cadre de compétitions en programmation. Ce jeu de données est largement utilisé en imagerie médicale [18].

Ce jeu de données contient un échantillon de 10 015 images réparties en sept classes différentes correspondant à l'état des grains de beauté. Comme cette recherche se concentre sur les grains de beauté de type mélanome et ceux qui sont bénins, nous avons décidé de nous limiter à deux classes. De plus, contrairement aux données initiales, les données utilisées ont déjà été prétraitées. Cela signifie que certains détails indésirables, comme les poils, ont été retirés et que les images ont été converties en niveaux de gris afin de mieux contraster le grain de beauté avec le reste de la peau.

3.1 Méthodes de prétraitement

Afin de rendre les données exploitables, plusieurs techniques de prétraitement ont été appliquées pour faciliter l'extraction des contours. Pour ce faire, nous avons utilisé deux étapes consécutives permettant d'isoler efficacement la forme des grains de beauté, soit : (1) le retrait des poils et (2) la combinaison optimale des canaux de couleurs pour la segmentation [4].

3.1.1 Retrait des poils

Sur la grande majorité des images, des poils traversant les grains de beauté peuvent être observés. Par conséquent, l'extraction des contours risque de prendre en compte ces poils, ce qui introduirait un biais nuisible à l'expérimentation. En effet, les poils ne fournissent aucune information pertinente sur l'état d'un grain de beauté. Ainsi, deux algorithmes sont utilisés afin de les retirer.

Dans un premier temps, l'algorithme du chapeau haut-de-forme est employé afin d'isoler les éléments clairs sur un fond sombre [17]. Dans un deuxième temps, une méthode de retouche d'image est utilisée afin de reconstruire les régions manquantes, détériorées ou masquées, de manière à rendre la modification imperceptible à l'œil nu [12].

3.1.2 Sélection de canaux pour la segmentation

Cette approche utilise des images binaires afin de faciliter la segmentation de la lésion et d'obtenir une représentation plus simple de sa forme. Une image binaire est composée de pixels auxquels sont attribuées deux valeurs, généralement (0) et (1), correspondant respectivement à l'arrière-plan et à l'objet d'intérêt. Cette représentation permet ainsi de distinguer plus facilement la lésion du reste de l'image. Avant la binarisation, l'image originale peut être décomposée en différents canaux de couleur, chacun correspondant à une composante du modèle colorimétrique utilisé. La sélection du canal dans lequel la lésion présente le contraste le plus marqué avec son environnement facilite alors son identification et peut améliorer la qualité de la segmentation. Une fois l'image binarisée, les pixels appartenant à la lésion peuvent être regroupés afin d'obtenir sa région d'intérêt, puis son contour peut être extrait. Ce contour constitue ensuite la représentation de la forme utilisée dans les étapes d'analyse subséquentes. Cette étape est particulièrement importante, puisque les erreurs introduites lors de la segmentation peuvent se répercuter sur l'extraction du contour et, par conséquent, sur l'analyse statistique de la forme. Par la suite, une nouvelle matrice est créée en combinant les différents canaux, puis une binarisation est effectuée afin d'obtenir des contours plus nets.

Finalement, une fois le prétraitement terminé, l'algorithme de Marching Squares est utilisé afin d'extraire le contour d'une image binaire [3]. Dans la figure 1, un exemple de contour extrait à l'aide de l'algorithme de Marching Squares est présenté.

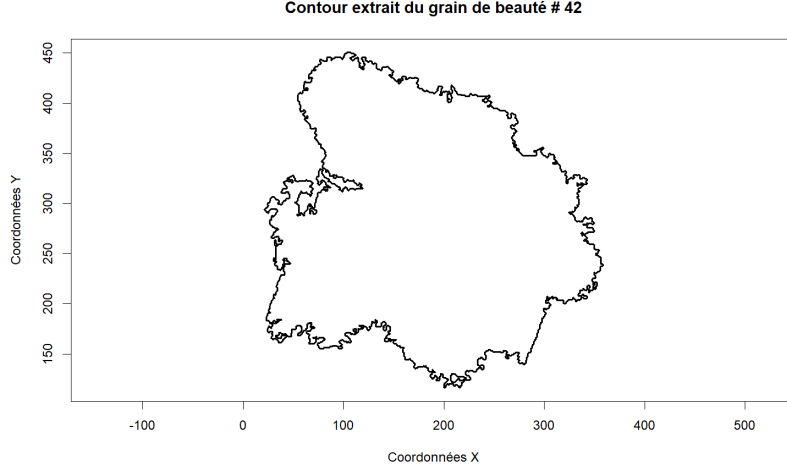


FIGURE 1 – Contour obtenu après l'extraction des frontières du grain de beauté 42.

4 Méthodologie

4.1 Fondements mathématiques de l'ADF

Les données fonctionnelles sont des observations décrivant l'évolution d'un individu, d'un système ou d'un objet au cours du temps. Chaque observation est décrite par une fonction, $X_i(t)$, qui dépend du temps t . Comme les fonctions évoluent continuellement dans le temps, certaines tâches classiques de l'analyse de données doivent être redéfinies. Par exemple, la régression linéaire pour des données discrètes est représentée de la façon suivante :

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i + \epsilon_i$$

Cependant, pour des données continues, ce même modèle est représenté de la manière suivante :

$$Y = \beta_0 + \int X(t)\beta(t)dt + \epsilon(t)$$

Comme les données sont capturées de manière discrète, la première étape consiste à appliquer un lissage afin de transformer ces observations en fonctions continues en utilisant, par exemple, une base de Fourier. En effet, la base de Fourier permet de réduire le bruit et de représenter plus facilement des trajectoires bidimensionnelles cycliques, sans nécessairement qu'elles soient parfaitement circulaires [16].

Les variables fonctionnelles peuvent être représentées sous la forme d'une combinaison de fonctions de base :

$$X(t) = \sum_{k=1}^K c_k \phi_k(t)$$

où $\phi_k(t)$ représente une fonction de base.

Plusieurs bases peuvent être utilisées, notamment la base de Fourier, qui repose sur des fonctions trigonométriques, ainsi que la base de B-splines. Dans le cadre de notre analyse de formes, qui sont représentées par des contours fermés, l'utilisation de la base de Fourier est privilégiée puisque ses fonctions sont intrinsèquement périodiques avec une période de 2π .

4.2 Représentation fonctionnelle des contours

Pour représenter les contours, qui se situent dans un espace bidimensionnel, ceux-ci sont décrits de la façon suivante :

$$C(t) = \begin{pmatrix} X(t) \\ Y(t) \end{pmatrix}, \quad t \in [0, 1]$$

Puisque les contours sont fermés, les coordonnées peuvent être représentées par :

$$\begin{aligned} x(t) &= a_0 + \sum_{k=1}^K (a_k \cos(2\pi kt) + b_k \sin(2\pi kt)) \\ y(t) &= c_0 + \sum_{k=1}^K (c_k \cos(2\pi kt) + d_k \sin(2\pi kt)) \end{aligned}$$

Les prédictors obtenus sont alors :

$$(a_k, b_k, c_k, d_k)$$

Ainsi, pour chaque contour i , un vecteur de caractéristiques est obtenu :

$$Z_i = (a_1, b_1, c_1, d_1, \dots, a_K, b_K, c_K, d_K)$$

Ces caractéristiques correspondent aux parties réelles et imaginaires des coefficients de Fourier associés aux coordonnées x et y dans la représentation complexe du contour [10].

4.3 Modèle de déformation des contours

Une version complexe du contour de référence $\tilde{C}$ peut être représentée par l'équation suivante :

$$C(t) = \rho O_\theta \tilde{C} \circ \gamma(t) + T, \quad t \in [0, 1]$$

où $C = (X, Y)^T$ est une fonction bivariée qui prend des valeurs dans l'espace des fonctions intégrables carrées $L^2([0, 1], \mathbb{R}^2)$ et

$$T = \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

est un vecteur contenant la translation selon les différentes coordonnées.

Ce modèle met en contexte quatre déformations géométriques. La translation (T) est un vecteur aléatoire dans $\mathbb{R}^2$ qui déplace la courbe dans le plan. La rotation (O) est une matrice de rotation basée sur un angle aléatoire $\theta \in [0, 2\pi]$ dans le but de faire tourner les contours. L'échelle (ρ) est un scalaire aléatoire positif qui grossit ou rapetisse la courbe. Finalement, la reparamétrisation (γ) est une fonction qui modifie la vitesse de parcours de la courbe [14].

Pour que ce modèle fonctionne de façon mathématique, plusieurs hypothèses et contraintes sont mises en avant. Premièrement, les courbes doivent être modélisées dans l'espace de Hilbert ($\mathcal{H}$), qui est un espace de fonctions de carré intégrable L^2 . Il est accompagné d'un produit scalaire standard, c'est-à-dire l'intégrale d'un produit de coordonnées. De plus, comme les grains de beauté sont des courbes lisses et fermées sur l'intervalle $[0, 1]$, on est obligé d'ajouter une contrainte de sorte que $C(0) = C(1)$. Aussi, la forme est normalisée, c'est-à-dire qu'elle appartient à la sphère unité $\left(\|\tilde{C}\|_{\mathcal{H}} = 1\right)$ dans le but de fixer l'échelle de référence, et centrée dans le but que l'intégrale des coordonnées dans $[0, 1]$ donne 0 [14].

4.4 Méthodes algorithmiques

4.4.1 Extraction des contours par Marching Squares

L'algorithme de Marching Squares est utilisé pour extraire les contours de formes à partir d'images. D'un point de vue mathématique, cet algorithme cherche à déterminer des courbes particulières situées sur la surface $z = f(x, y)$. Ces courbes sont appelées des courbes de niveau et sont obtenues par l'intersection entre la surface et un plan horizontal $z = c$, où $c \in \mathbb{R}$. Ces courbes dans le plan xy correspondent aux ensembles de niveaux de la fonction d'origine f [3].

4.4.2 Lissage des données

Le lissage permet de transformer des données discrètes en données continues afin qu'elles soient dérivables et représentables sous forme de courbes, tout en réduisant l'effet du bruit. Initialement, les données sont représentées de la façon suivante : $x(t_i)$, où t_i représente

le i -ième instant d'observation. Ensuite, à l'aide d'un lissage par bases, les données sont approximées par la représentation suivante :

$$x(t) = \sum_{k=1}^K c_k \phi_k(t)$$

Par la suite, les coefficients c_k sont choisis de manière à minimiser :

$$\sum_{i=1}^n (y - x(t_i))^2 + \lambda \int [D^2 x(t)]^2 dt$$

où D^2 représente la dérivée seconde de la donnée fonctionnelle. De manière précise, $D^2 x(t) = x''(t)$ mesure la variation de cette pente. Donc, nous pénalisons les variations importantes de la courbure [16].

4.4.3 Alignement des contours par ICF

L'algorithme *Iterative Closest Function* (ICF) sert à aligner deux courbes planaires fermées en considérant une rotation θ et un point de départ δ . Pour ce faire, l'algorithme cherche à minimiser de manière itérative la distance entre ces courbes jusqu'à atteindre un critère de convergence. L'estimation repose principalement sur deux paramètres : l'angle de rotation θ et la déformation du paramètre de parcours δ . Pour une valeur fixée de δ , le paramètre θ est obtenu en résolvant le problème de Procrustes optimal suivant :

$$\hat{\theta} = \arg \min_{\theta \in [0, 2\pi]} \|\mathbf{O} * \theta \boldsymbol{\mu} \circ \gamma * \delta - \mathbf{C}^*\|_{\mathcal{H}}^2$$

Par la suite, pour un paramètre θ fixé, le paramètre δ est estimé numériquement à l'aide d'une méthode de bisection. La répétition de ces étapes est effectuée jusqu'à la convergence de l'algorithme. Comme celui-ci dépend de l'initialisation de δ , il peut être très sensible au choix de la valeur initiale. C'est pourquoi plusieurs valeurs initiales de δ sont considérées afin de sélectionner la paire de paramètres $(\hat{\theta}, \hat{\delta})$ minimisant l'objectif principal [14].

4.4.4 Analyse fonctionnelle en composantes principales (FPCA)

Cette technique constitue une variante de l'analyse en composantes principales (ACP), puisqu'elle traite des données fonctionnelles, c'est-à-dire des courbes ou des formes. En projetant les données fonctionnelles dans un espace de dimension réduite, cette approche permet de mettre en évidence les principaux modes de variation, qui peuvent être difficiles à identifier dans un espace de grande dimension.

L'algorithme est présenté de la façon suivante :

1. Obtenir des courbes lisses à partir des données discrètes.

$$X_1(t), X_2(t), \dots, X_N(t)$$

2. Calculer la moyenne fonctionnelle.

$$\mu(t) = \frac{1}{N} \sum_{i=1}^N X_i(t)$$

3. Centrer les données.

$$X_i^*(t) = X_i(t) - \mu(t)$$

4. Calculer la fonction de covariance.

$$v(s, t) = \frac{1}{N} \sum_{i=1}^N X_i^*(s) X_i^*(t)$$

5. Application de la décomposition spectrale.

$$\int v(s, t) \epsilon(t) dt = \rho \epsilon(s)$$

où ρ correspond aux valeurs propres et $\epsilon(s)$ aux fonctions propres.

6. Projection sur un espace à basse dimension.

On fait comme en ACP avec les $\epsilon_1, \epsilon_2, \dots, \epsilon_{j-1}$, c'est-à-dire des fonctions propres associées aux plus grandes valeurs propres de l'opérateur de covariance. Les données sont ensuite projetées sur ces fonctions propres afin d'obtenir leurs composantes principales [16].

4.4.5 Classification par régression logistique pondérée

Finalement, pour effectuer la classification, nous utilisons une régression logistique pondérée dont l'équation est la suivante :

$$\log \left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \right) = \beta_0 + \beta_1 t_x + \beta_2 t_y + \beta_3 \theta + \beta_4 E + \sum_{k=1}^K \beta_{k+4} \xi_k$$

où :

- t_x, t_y : paramètres de translation du contour,
- θ : angle de rotation du contour,
- E_i : erreur de reconstruction par rapport au contour moyen,
- ξ_{ik} : scores associés aux k premières composantes principales fonctionnelles (FPCA),
- Y_i : classe de la lésion (bénigne ou mélanome) [9].

La régression logistique permet d'estimer la probabilité qu'une observation appartienne à la classe mélanome ($Y = 1$) en fonction des caractéristiques géométriques et des scores provenant de la FPCA.

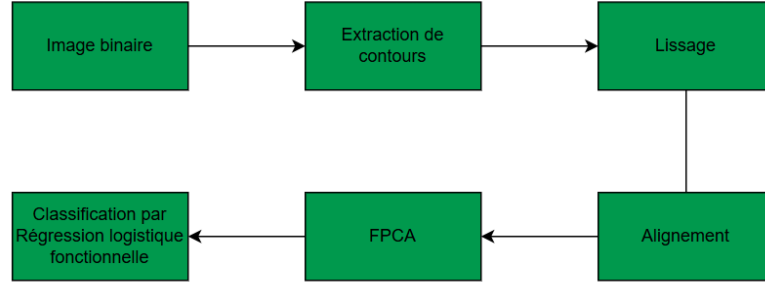


FIGURE 2 – Pipeline de classification des grains de beauté.

Un résumé du processus complet de l'approche proposé est résumé dans la figure 2.

4.5 Stratégies de validation

Afin de s'assurer que le modèle fonctionne correctement, le jeu de données est séparé en deux parties. Ainsi, 80% des données sont utilisées pour l'entraînement du modèle, tandis que 20% sont réservées au test afin d'évaluer la performance de la méthodologie proposée.

De plus, pour valider la méthode utilisée, une validation par simulation est appropriée, puisqu'elle permet d'évaluer l'approche dans des conditions où la vérité sur la forme est connue. En effet, selon différents scénarios, des contours fonctionnels sont générés à partir de fonctions connues, en ajoutant du bruit et en modifiant différents paramètres.

4.6 Métriques de validation

La méthode est évaluée à l'aide de différentes métriques de performance, soit l'exactitude, la précision, le score F1 et l'AUC. L'utilisation de plusieurs métriques permet d'obtenir une évaluation plus complète des performances du classificateur, puisqu'elles ne mesurent pas toutes le même aspect de la classification.

L'exactitude mesure la proportion totale de prédictions correctement classées, tandis que la précision indique la proportion de prédictions positives qui sont réellement positives. Le score F1 combine la précision et le rappel et constitue ainsi une mesure particulièrement pertinente lorsque les classes sont déséquilibrées, comme c'est le cas dans notre situation où les grains de beauté bénins sont beaucoup plus nombreux que les mélanomes. Enfin, l'AUC mesure la capacité du classificateur à distinguer les deux classes pour différents seuils de décision.

Ainsi, une méthode peut présenter une bonne exactitude tout en ayant de moins bonnes performances pour la détection des mélanomes, notamment en raison du déséquilibre

entre les classes. L'analyse conjointe de ces métriques permet donc de mieux évaluer la capacité du modèle à identifier correctement les mélanomes, plutôt que de se limiter à sa performance globale.

5 Résultats

La section suivante présente les résultats obtenus après l'application de la méthode de classification. La figure 3 présente les scores associés à chaque composante principale, tandis que le tableau 1 illustre les résultats des différentes métriques pour chaque grain de beauté de référence.

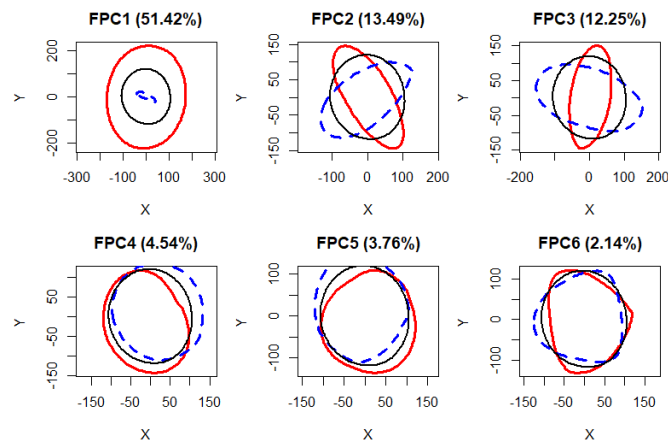


FIGURE 3 – Graphiques des composantes principales des contours de grains de beauté. Le premier contour est utilisé comme référence. Le contour noir représente la taille moyenne, le contour bleu la taille minimale et le contour rouge la taille maximale.

Grain de beauté de référence	AUC	F1	Exactitude
1	0.5966	0.2852	0.4852
2	0.5960	0.2830	0.7105
3	0.6015	0.2850	0.4557
4	0.6028	0.2844	0.3928
5	0.6052	0.2896	0.5687

TABLE 1 – Métriques de performance selon différents grains de beauté de référence en utilisant la Régression logistique fonctionnelle

5.1 Analyse des erreurs

À partir des résultats obtenus, nous pouvons remarquer que certaines métriques de performance présentent des valeurs faibles. Cela peut être expliqué par différentes sources d'erreurs introduites lors du processus d'analyse. Premièrement, dans le jeu de données

initial, plusieurs images contiennent du bruit ainsi que des variations d'éclairage. En effet, chaque grain de beauté possède un contour avec une forme unique. Par conséquent, certaines informations indésirables peuvent être prises en compte lors de l'analyse, alors qu'elles ne permettent pas de caractériser le type de grain de beauté. À l'inverse, certains contours peuvent présenter des imprécisions difficiles à extraire.

De plus, concernant les conditions d'éclairage, certaines images présentent des grains de beauté fortement éclairés, tandis que d'autres sont plus sombres. En conséquence, les contours sont plus ou moins visibles, ce qui peut rendre l'extraction des frontières moins précise pour certains grains de beauté.

Par ailleurs, en ce qui concerne les classes, un déséquilibre est présent entre la classe des mélanomes et celle des grains de beauté bénins. Environ 14.18% des grains de beauté correspondent à des mélanomes, tandis que 85.82% correspondent à des lésions bénignes [18]. Comme les grains de beauté bénins représentent la classe majoritaire, le modèle de classification peut avoir tendance à favoriser cette classe afin de minimiser l'erreur globale. Cela entraîne un biais vers la classe majoritaire et peut réduire la capacité du modèle à détecter les observations appartenant à la classe minoritaire [1].

6 Discussion

À partir des résultats obtenus, nous pouvons remarquer que l'utilisation de la base de Fourier est appropriée, puisque les contours des grains de beauté sont des formes fermées, lisses et périodiques. À l'inverse, la base de B-splines est moins adaptée dans ce contexte, car elle ne garantit pas naturellement la périodicité et nécessite l'ajout de contraintes supplémentaires pour fermer correctement les contours. De plus, lorsque le nombre de fonctions de base est faible, la représentation par Fourier peut être moins sensible aux détails locaux, comme les irrégularités présentes sur les frontières, ce qui entraîne un lissage global.

Selon les résultats des métriques présentés dans le tableau 1, l'AUC obtient des valeurs autour de 0.60, le score F1 autour de 0.30, tandis que l'exactitude varie entre 0.39 et 0.72. Ces résultats indiquent que le modèle possède une capacité de discrimination légèrement supérieure à une classification aléatoire. Cette performance peut être expliquée par le fait que la régression logistique prend en compte les résidus, le paramètre de rotation θ ainsi que les scores issus de la FPCA. En effet, l'alignement permet de réduire les différences d'orientation entre les contours en effectuant des rotations par rapport à une forme de référence.

Cependant, les performances restent limitées en raison de plusieurs facteurs. Premièrement, la simplicité du modèle ne permet pas de capturer toute la complexité biologique des lésions. Les informations utilisées reposent principalement sur la géométrie des contours, alors que d'autres caractéristiques importantes comme la texture, la couleur ou

la structure interne pourraient contribuer à la classification. Deuxièmement, le déséquilibre entre les classes influence les performances. En effet, moins de 15% des observations appartiennent à la classe des mélanomes, tandis que la majorité correspond aux grains de beauté bénins. Ainsi, la classe majoritaire peut influencer davantage le modèle et réduire sa capacité à détecter les observations minoritaires. De plus, un mauvais alignement peut modifier les coefficients de Fourier ou les composantes principales, ce qui pourrait affecter la classification [16].

Concernant les composantes principales, la somme des scores associés aux différentes composantes explique une proportion de variance comprise entre 87% et 88%. Les premières composantes représentent principalement des variations globales des contours, telles que l'échelle ou l'orientation, tandis que les dernières composantes capturent davantage les déformations locales et les variations de forme [20]. Certaines variations de type cardioïde peuvent notamment être observées, où une partie du contour devient plus saillante tandis que la partie opposée est comprimée. Toutefois, même si la FPCA permet de réduire la dimensionnalité tout en conservant une grande partie de la variance, ces composantes ne suffisent pas à représenter l'ensemble de la complexité morphologique associée aux lésions cutanées.

En comparant les résultats obtenus avec ceux présentés dans l'étude d'Andréa [4], nous observons des performances similaires pour l'AUC, mais des résultats inférieurs pour les autres métriques. Cette différence peut être expliquée par l'utilisation des réseaux CNN, qui sont capables d'extraire automatiquement des caractéristiques pertinentes directement à partir des images [21]. De plus, l'étude utilise l'exactitude équilibrée comme métrique, ce qui permet de mieux prendre en compte le déséquilibre entre les classes. Contrairement à l'exactitude classique, cette métrique attribue un poids équivalent aux différentes catégories et fournit ainsi une évaluation plus représentative des performances du modèle.

Globalement, les résultats obtenus sont partiellement satisfaisants. Bien que la régression logistique fonctionnelle permette une certaine distinction entre les mélanomes et les lésions bénignes, les informations géométriques provenant uniquement des contours ne semblent pas suffisantes pour capturer toutes les différences morphologiques entre les deux classes.

7 Conclusion

7.1 Synthèse des résultats

Cette étude avait pour objectif de montrer qu'une représentation fonctionnelle des données permettait de différencier les grains de beauté bénins de ceux correspondant à des mélanomes. La méthode proposée repose sur une représentation fonctionnelle des contours, en tenant compte de l'alignement obtenu à l'aide de l'algorithme de Marching Squares, suivi

d’une réduction de dimensionnalité par analyse en composantes principales fonctionnelle (FPCA) et d’une étape de classification utilisant une régression logistique pondérée. Les résultats obtenus concernent les scores des composantes principales, permettant d’analyser leur proportion de variabilité expliquée ainsi que leurs modes de variation, en plus d’évaluer les performances du modèle dans la classification des grains de beauté.

Les performances obtenues pour différentes formes de référence utilisées lors de l’alignement montrent une faible sensibilité du modèle au choix de la référence, avec des valeurs d’AUC variant entre 0.596 et 0.605. Toutefois, les performances globales demeurent limitées, ce qui suggère que l’information géométrique provenant uniquement du contour n’est pas suffisante pour distinguer efficacement les lésions mélanocytaires bénignes et malignes.

7.2 Limites et défis

Comme nous avons pu le constater, la base de Fourier est particulièrement adaptée aux données périodiques et aux fonctions continues. Bien que son lissage soit efficace et que ses fonctions soient orthogonales, cette représentation possède certaines limites [16]. En effet, lorsque des irrégularités sont présentes, elle peut avoir de la difficulté à capturer les variations locales sans modifier la structure globale des contours. Par exemple, la présence de poils traversant les grains de beauté peut influencer l’extraction des contours, faisant en sorte que le lissage intègre des informations indésirables.

De plus, notre approche repose uniquement sur les informations géométriques des contours et ne considère pas d’autres caractéristiques importantes, comme la taille et la couleur. Certains mélanomes peuvent ainsi présenter des contours similaires à ceux de lésions bénignes, ce qui complique leur classification par la régression logistique fonctionnelle. Toutefois, certaines caractéristiques comme une taille plus importante ou des variations de couleurs peuvent également être pertinentes [7].

Enfin, l’algorithme de Marching Squares peut être sensible au bruit et aux structures complexes. Dans le cas des grains de beauté, certaines petites irrégularités peuvent être ignorées lors de l’extraction des contours, tandis que l’algorithme peut aussi générer des boucles artificielles ou des incohérences topologiques.

7.3 Perspectives futures

Même si la base de Fourier est particulièrement adaptée aux contours lisses et fermés, l’utilisation d’autres bases demeure tout à fait possible. En effet, la base de B-splines peut être envisagée, puisqu’elle s’adapte très bien aux tendances complexes, aux variations rapides et aux structures locales. De plus, elle offre une grande flexibilité concernant la résolution temporelle ou spatiale grâce au choix des nœuds [10]. Toutefois, son principal

désavantage est qu'il faut imposer une condition de périodicité afin de l'appliquer à des contours fermés :

$$\begin{cases} X(0) = X(1), \\ Y(0) = Y(1), \\ C(0) = C(1). \end{cases}$$

La base de B-splines ne tient pas naturellement compte de la périodicité, à moins d'imposer une condition explicite. De plus, la base d'ondelettes aurait également pu être utilisée. Elle repose sur une fonction mère qui est dilatée et déplacée le long du domaine [19]. Elle est adaptée aux contours présentant à la fois des régions lisses et des variations brusques, car elle permet de capturer des discontinuités tout en limitant l'effet de Gibbs, c'est-à-dire un comportement oscillatoire indésirable apparaissant lors de l'approximation d'une fonction discontinue [8]. Cependant, tout comme la base de B-splines, elle n'assure pas naturellement la périodicité. Il faudrait donc imposer une condition supplémentaire pour l'appliquer à des contours fermés.

De plus, comme métrique d'évaluation, l'exactitude équilibrée pourrait être utilisée. Comme mentionné dans la discussion, cette métrique donne un poids égal aux différentes classes malgré un fort déséquilibre entre celles-ci. En effet, pour le premier contour comme référence, nous obtenons un résultat de 0.586 signifiant que le poids égal a été pris en compte. Ainsi, le modèle serait moins susceptible de favoriser la classe des bénins, qui est majoritaire.

Code

Ce projet s'est basé avec le langage R pour implémenter différents algorithmes et pour l'utilisation d'ADF. Le lien du projet se trouve ici : https://github.com/Kevin-Wang78/Stage_FDA.

Remerciements

Les auteurs voudraient remercier le support financier de STATQAM. Les auteurs voudraient aussi remercier Louis-Pierre Ménard pour ses suggestions.

Références

- [1] Ufuk Beyaztas and Han Lin Shang. Robust functional linear regression models. *R J.*, 15(1) :212–233, 2023.
- [2] Wilhelm Burger and Mark J Burge. Fourier shape descriptors. In *Principles of Digital Image Processing : Advanced Methods*, pages 169–227. Springer, 2013.
- [3] Quasar Chunawala. Drawing shapes with marching squares. <https://www.baeldung.com/cs/marching-squares>, year = 2023,.
- [4] Andréa Davila, Issam-Ali Moindjié, and Cédric Beaulac. Comparing shape-based and pixel-based approaches for melanoma detection. Technical report, Université du Québec à Montréal, February 2026.
- [5] Manuel Febrero-Bande and Manuel Oviedo De La Fuente. Statistical computing in functional data analysis : The r package fda. usc. *Journal of statistical Software*, 51 :1–28, 2012.
- [6] Joao B Florindo and Odemir M Bruno. Closed contour fractal dimension estimation by the fourier transform. *Chaos, Solitons & Fractals*, 44(10) :851–861, 2011.
- [7] Meenhard Herlyn. *Molecular and Cellular Biology of Melanoma*. R.G. Landes Company Austin, Philadelphie, 1993.
- [8] Susan E Kelly. Gibbs phenomenon for wavelets. *Applied and Computational Harmonic Analysis*, 3(1) :72–81, 1996.
- [9] David G Kleinbaum and Mitchel Klein. *Logistic regression : a self-learning text*. Springer, 2002.
- [10] Frank P. Kuhl and Charles Giardina. Parametrization of closed curves and surfaces. In *Elliptic Fourier Features of a Closed Contour*. Academic Press, 1982.
- [11] Robert Loiselle and Raynald Côté. *Guide de rédaction et de présentation d’un texte scientifique*. Université du Québec à Chicoutimi, Module de biologie., 1980.
- [12] Xinyue Luo, Yu Chen, Jin Cheng, and Zhaoyi Sun. A mechanism-based method for image inpainting. *Applied Mathematics Letters*, 154 :109113, 2024.
- [13] Benoit B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Company, New York, 1982.
- [14] Issam-Ali Moindjié, Cédric Beaulac, and Marie-Hélène Descary. A functional approach for curve alignment and shape analysis. *arXiv preprint arXiv :2503.05632*, 2025.

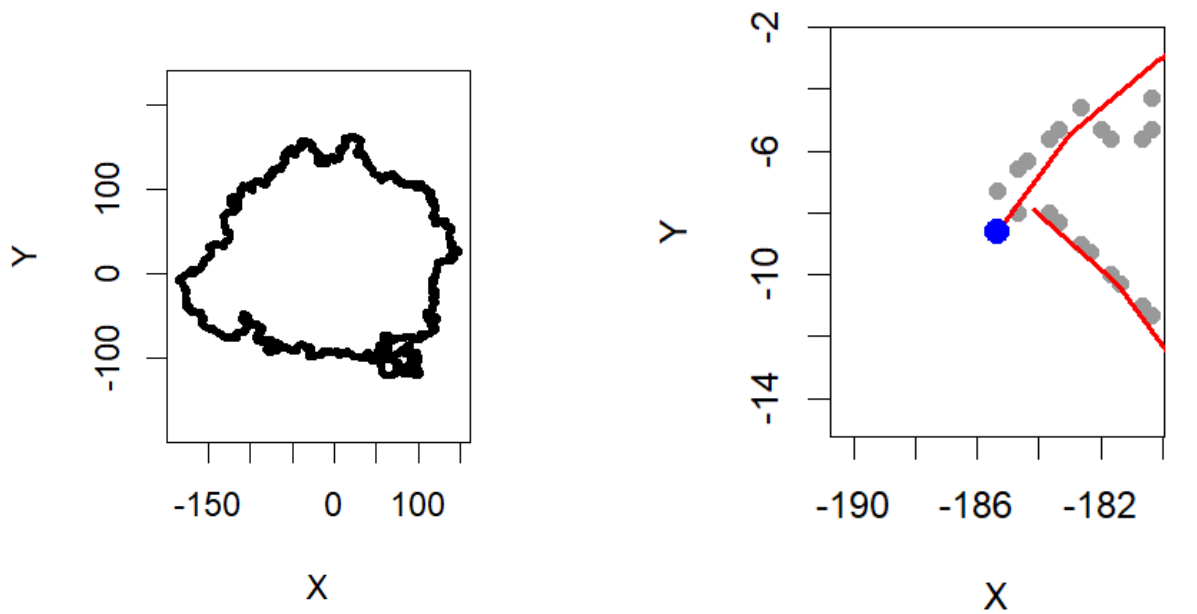
- [15] Hoda Naseri and Ali A Safaei. Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning : a systematic literature review. *BMC cancer*, 25(1) :75, 2025.
- [16] James O Ramsay and Bernard W Silverman. *Functional data analysis*. Springer, 1997.
- [17] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982.
- [18] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1) :180161, 2018.
- [19] David F Walnut. *An introduction to wavelet analysis*. Springer Science & Business Media, 2013.
- [20] Guy Wolf. Notes de cours : Fondements théoriques en science de données (stt 3795). Notes de cours, Département de mathématiques et de statistique, Université de Montréal, 2026.
- [21] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23) :495, 2017.

A Figures supplémentaires

Cette section montre des figures supplémentaires obtenus par différents algorithmes.



FIGURE 4 – Image du grain de beauté 42 extraite par transformation binaire



(a) Image du contour d'un grain de beauté lissé avec une base de B-splines

(b) Image précise du contour de ce grain de beauté vue de façon miniature

FIGURE 5 – Illustrations montrant que l'utilisation de la base de B-splines est erronée pour les grains de beauté

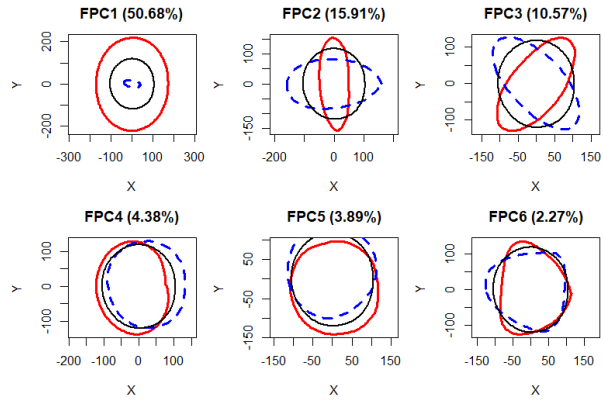


FIGURE 6 – Graphique des composantes principales avec le pourcentage de variabilité expliquée où le deuxième grain de beauté est la référence

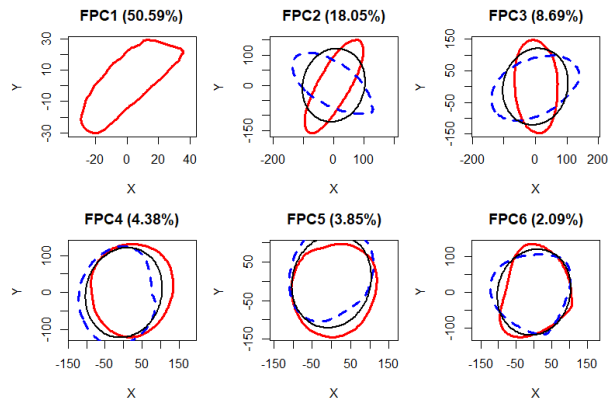


FIGURE 7 – Graphique des composantes principales avec le pourcentage de variabilité expliquée où le troisième grain de beauté est la référence

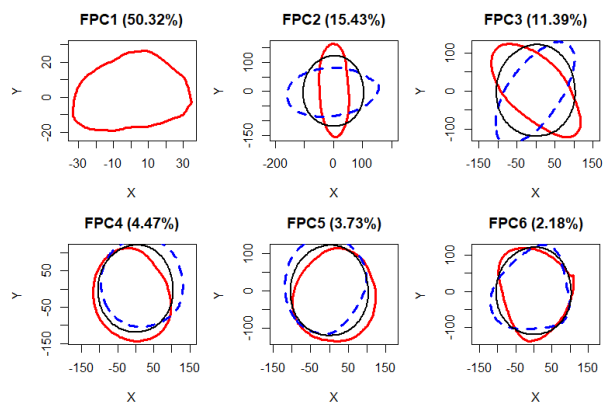


FIGURE 8 – Graphique des composantes principales avec le pourcentage de variabilité expliquée où le quatrième grain de beauté est la référence

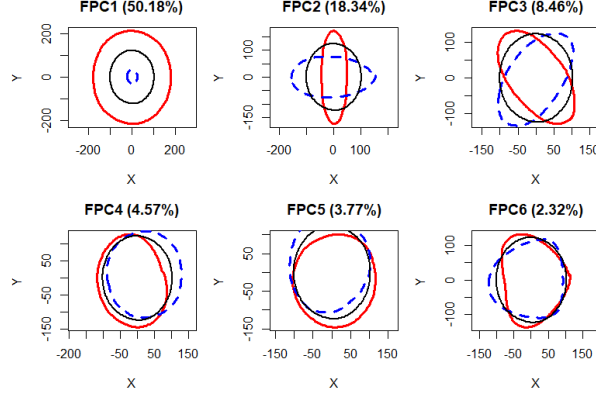


FIGURE 9 – Graphique des composantes principales avec le pourcentage de variabilité expliquée où le cinquième grain de beauté est la référence

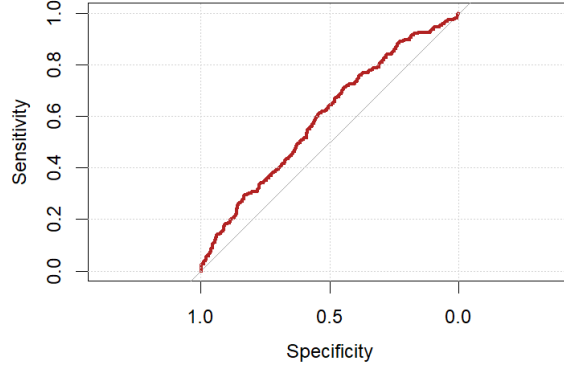


FIGURE 10 – Courbe ROC illustrant le compromis entre la sensibilité et la spécificité pour le premier grain de beauté comme référence

B Preuves des résultats

La propriété suivante est inspirée des démonstrations présentées dans [2] et dans [6]. Une preuve détaillée adaptée au contexte des grains de beauté est donnée ci-dessous.

Première mathématique : Montrer que pour des contours lisses, fermés et périodiques, la base de Fourier est toujours plus appréciée que celle de B-spline.

Situation initiale

Soit un contour fermé de type :

$$C(t) = \begin{pmatrix} X(t) \\ Y(t) \end{pmatrix}, t \in [0, 1]$$

Comme le contour d'un grain de beauté est fermé, alors, $C(0) = C(1)$ par définition. Cela veut dire en d'autres termes que $X(0) = X(1)$ et que $Y(0) = Y(1)$ signifiant que ce sont des fonctions périodiques.

Théorème de Fourier

Nous savons que les fonctions :

$$\{1, \cos(2\pi kt), \sin(2\pi kt)\}_{k \geq 1}$$

forment une base orthogonale et de norme unitaire. En conséquence,

$$\begin{cases} x(t) = a_0 + \sum_{k \geq 1} (a_k \cos(2\pi kt) + b_k \sin(2\pi kt)) \\ y(t) = c_0 + \sum_{k \geq 1} (c_k \cos(2\pi kt) + d_k \sin(2\pi kt)) \end{cases}$$

Approximation optimale (Parseval)

Maintenant, supposons qu'on garde seulement K fonctions harmoniques parmi l'infinité de fonctions.

L'approximation est ainsi :

$$x_K(t) = a_0 + \sum_{k=1}^K (a_k \cos(2\pi kt) + b_k \sin(2\pi kt))$$

Le théorème de Parseval énonce que :

$$\|x - x_K\|_{L^2}^2 = \sum_{k > K} (a_k^2 + b_k^2).$$

C'est la meilleure approximation possible parmi toutes les combinaisons linéaires avec les $2K + 1$ premiers modes de Fourier.

En effet,

$$x_K = \operatorname{argmin}_{v \in V_K} \|x - v\|_{L^2},$$

avec le fait que

$$V_K = \operatorname{span}\{1, \cos, \sin, \dots\}.$$

Décroissance des coefficients

Supposons que :

$$x \in H^s([0, 1])$$

Donc,

$$|a_k|, |b_k| = O(k^{-s})$$

En conséquence,

$$\|x - x_K\|^2 = \sum_{k>K} O(k^{-2s}) = O(K^{-(2s-1)})$$

Comme on peut le remarquer, les coefficients décroissent rapidement.

Cas de la base de B-spline

Pour le cas de cette base, une approximation s'écrit :

$$x(t) = \sum_{i=1}^m c_i B_i(t)$$

où B_i sont les fonctions de cette base. Comme les B-splines sont locales, alors, chaque base est uniquement non nulle sur un petit intervalle. Le seul problème est que les bases ne sont pas orthogonales et la matrice de Gram n'est pas diagonale.

Conditions périodiques

Pour que la base de B-spline soit utilisée pour les contours lisses, fermés et périodiques, il faut qu'on impose obligatoirement une condition de la forme :

$$x(0) = x(1)$$

et même chose pour ses dérivées première et seconde. Donc, il faudra construire des B-splines périodiques.

C Métriques de performance

Voici les définitions de métriques de performance présentes dans la section 3' :

- Vrai positif := VP,
- Vrai négatif := VN,
- Faux positif := FP,
- Faux négatif := FN,
- True Positive Rate := $TPR = \frac{VP}{VP+FN}$,
- True Negative Rate := $TNR = \frac{TN}{FP+VN}$,
- False Positive Rate := $FPR = \frac{FP}{FP+VN}$,
- Area Under the Curve := $AUC = \int_0^1 TPR(FPR)d(FPR)$, où $d(FPR)$ est l'élément différentiel de FPR
- F1 := $\frac{2VP}{2VP+FP+FN}$,
- Exactitude := $\frac{VP+VN}{VP+VN+FP+FN}$