الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de 8 Mai 1945 – Guelma -

Faculté des Mathématiques, d'Informatique et des Sciences de la matière

Département d'Informatique

Mémoire de Fin d'études Master

Filière : Informatique

Option : Ingénierie des médias

Thème :

Une Approche génétique pour la sélection de caractéristiques des masses mammographiques.

Encadré par :

Mr.FERKOUS Chokri

Présenté par :

ZIADA Samir

BOUREGBA MohammedAli

Juin 2017

Remerciements

Au nom de dieu le clément, le miséricordieux

nous remercie Dieu tout Puissant de nous avoir permis de mener à terme ce mini projet qui est pour nous le point de départ d'une merveilleuse aventure, celle de la recherche, source de remise en cause permanent et de perfectionnement perpétuelle.

Qu'il nous soit permis de rendre un vibrant hommage à notre encadreur, Mr. FERKOUS CHOKRI pour avoir bien voulu superviser ce modeste travail et donné de son temps et de son intelligence à la réussite de ce projet qui pour nous représente un modèle de réussite et une source de motivation permanente, pour sa disponibilité, et son sens aigu de l'humanisme pédagogique.

Nous tenons à remercier l'ensemble du personnel du département d'informatique pour son aide fructueuse.

Mes remerciements vont également à l'endroit de mes enseignants, leur disponibilité et leurs précieux enseignements mon été d'une grande utilité.

Nous exprime notre gratitude à tous les consultants et internautes rencontrés lors des recherches effectuées et qui ont accepté de répondre à nos questions avec gentillesse.

Les mots me manquent pour exprimer ma profonde reconnaissance à ma tendre famille dont l'amour, la patience et le sacrifice s'inscrivent à chaque page de ce document.

Sans oublier tous nos amis qui nous ont aidés et encouragés même par leurs pensées, qui nous aiment et que nous aimons

Enfin je remercie les membres du jury qui ont bien voulu accepter, et ce nonobstant, leur lourdes et exaltantes responsabilités pour procéder à l'évaluation de ce modeste travail.

Dédicace

Avant tout je tiens à remercier notre grand seigneur qui m'a donné la force et le courage pour atteindre ce niveau.

Je dédie ce modeste travail à celle qui m'a donné la vie, Aux êtres les plus chers à mon cœur, MA Mère, qui a toujours cru en moi et qui m'a assurée un soutien inconditionnel sans lequel je n'aurais jamais pu terminer mes années d'études pour ses conseils et son encouragement.

A mon père, pour m'avoir donné la possibilité de faire ce que Je voulais et pour son soutien sans faille tout au long de mes études et à son affection, sa patience et sa compréhension.

A mes chères sœurs : Naziha,Djezira, chers frères :Hacene,Mehdi,Salim.

Je remercie ma belle famille, surtout mes tantes et mes oncles, mes cousins et mes cousines.

A mes meilleurs amis qui ont toujours été là pour moi et avec qui je partage des moments très agréables.

Un grand merci à tous ceux qui m'aiment.

A tous ceux que j'aime.

Samir.

Dédicace

Avant tout je tiens à remercier notre grand seigneur qui m'a donné la force et le courage pour atteindre ce niveau.

Je dédie ce modeste travail à celle qui m'a donné la vie, Aux êtres les plus chers à mon cœur, MA Mère, qui a toujours cru en moi et qui m'a assurée un soutien inconditionnel sans lequel je n'aurais jamais pu terminer mes années d'études pour ses conseils et son encouragement.

A mon père, pour m'avoir donné la possibilité de faire ce que Je voulais et pour son soutien sans faille tout au long de mes études et à son affection, sa patience et sa compréhension.

A mes chères sœurs, chers frères.

Je remercie ma belle famille, surtout mes tantes et mes oncles, mes cousins et mes cousines.

A mes meilleurs amis qui ont toujours été là pour moi et avec qui je partage des moments très agréables.

Un grand merci à tous ceux qui m'aiment.

A tous ceux que j'aime.

Mohammed Ali.

Résumé

Le cancer du sein représente l'un des enjeux prépondérants dans le domaine de la santé publique puisqu'il touche une grande population féminine et constitue à lui seul le plus grand pourcentage de mortalité chez les femmes. Cependant la détection des tumeurs à une étape précoce augmentera considérablement les chances de guérison des patientes. Il apparait indispensable de développer de nouvelles méthodes ou approches pour le dépistage du cancer du sein.

Dans ce mémoire nous proposons un système pour la sélection des caractéristiques des masses mammographiques, ce système s'inspire globalement de l'approche génétique lors de l'examen radiologique en se basant sur le concept des algorithmes génétiques qui permet à partir des opérateurs génétiques la sélection des caractéristiques pertinentes.

Nous avons validé notre système sur un échantillon de 59 caractéristiques calculées à partir de la zone d'intérêt, et les résultats obtenus sont très encourageants.

Mots clés :

La sélection de caractéristiques, algorithmes génétiques, mammographie.

Sommaire

Table des Figures

Liste des tableaux

Liste des abréviations

ACR	American College of Radiology
ANN	Artificial Neural Networks
AMDI	Indexed Atlas of Digital Mammograms
AG	Algorithme génétique
BI-RADS	Breast Imaging-Reporting And Data System
CAD	Computer Aided Diagnosis
DAOx	Diagnostic assisté par ordinateur
DDSM	Digital Database for Screening Mommography
ECG	électrocardiographie
EEG	électroencéphalogramme
FN	Faux négatifs
FP	Faux positifs
IRM	Imagerie par Résonance Magnétique
MCs	Micro Calcifications
MIAS	Mammographic Image Analysis Society
SDA	Systèmes de Dépistage Automatique
PMC	Perceptron multi-couches
LDE	Large Différences Emphasis
LRN	La Longeur Radiale Normalisé
RDF	Reconnaissance de forme
WDBC	Wisconsin Diagnosis Breast Cancer

Introduction Générale

Le cancer du sein représente l'un des enjeux prépondérants dans le domaine de la santé publique. Il est constitue, à travers toute la planète, la première cause de décès chez les femmes. Avec d'un million de nouveaux cas diagnostiqués chaque année dans le monde et plus que 400000 vont mourir de cette pathologie, soit 30 % des nouveaux cas de cancers dans les pays industrialisés et 14% dans les pays en voie de développement [33], le cancer du sein est le cancer le plus prévalent au monde. En Algérie, chaque année, environ 7500 cas de cancer du sein sont enregistrés, sachant que ce type de maladie en Algérie vient en tête des tumeurs malignes chez la femme et constitue la première cause de mortalité chez les jeunes femmes, avec environ 3500 décès enregistrés chaque année [34]. Le cancer du sein représente près de 50 % des cancers gynécologiques chez la femme, au cours de ces 15 dernières années où l'incidence du cancer du sein a été multipliée par 3. Du fait de son diagnostic tardif, il en résulte souvent un traitement lourd, mutilant et coûteux qui s'accompagne d'un taux de mortalité élevé. L'incidence du cancer du sein reste croissante en Algérie et il n'existe toujours pas de programmes de dépistage organisés à l'instar de nos voisins Maghrébins ou Européens.

Dans le but d'éviter des traitements lourds et de réduire la morbidité et la mortalité par le cancer du sein, une détection précoce est nécessaire motivant ainsi des campagnes de dépistage chez les femmes à partir d'un certain âge variant entre 40 et 50 ans.

Les techniques d'imagerie médicale sont multiples basées sur différents types de rayonnements. Le succès de ces techniques s'explique principalement par le fait que l'information médicale qu'elles apportent est de plus en plus précise pour une nocivité de l'examen de plus en plus faible. Notons que la mammographie est la technique la plus performante et la plus reproductible pour le diagnostic précoce du cancer du sein. Dans l'immense majorité des cas, elle est le premier examen d'imagerie radiologique consacré à la détection des pathologies mammaires. Cet examen s'effectue avec un appareil dédie uniquement à cet usage c'est le mammographe. Cet appareillage utilise les rayons X pour produire des images de hautes résolutions du sein.

L'interprétation des clichés en un temps limité et la détection des pathologies est une tâche très difficile par un examen mammographique, cela nécessite une double lecture, à travers l'utilisation de systèmes de dépistage automatique (SDA) afin d'améliorer les performances du lecteur.

Un système de dépistage automatique (SDA) typique comporte essentiellement quatre étapes principales : la segmentation, l'extraction de caractéristiques, la sélection de caractéristiques et la classification. La phase de sélection de caractéristique est l'étape qui nous intéresse le plus à case de son influence sur l'étape de classification. Pour cela nous avons proposé une approche génétique pour cette tâche.

Les algorithmes génétiques font partie de la famille des algorithmes dits « évolutifs » fondés sur les mécanismes de la sélection naturelle et de la génétique. Leur fonctionnement extrêmement simple. On part d'une population de solutions potentielles (chromosomes) initiales arbitrairement choisi. Chaque chromosome est évalué à travers une performance relative (fitness) ce qui permet de quantifier sa qualité. Sur la base de ces performances nous pouvons créer une nouvelle population de solution potentielle en utilisant des opérateurs évolutionnaire: la sélection, le croisement, la mutation.

Dans ce projet, nous proposons une approche génétique pour la sélection de caractéristiques à partir de plusieurs descripteurs calculés de l'image mammographique pour la classification des masses en deux grandes catégories (Maligne et bénigne).

Ce mémoire est composé de quatre chapitres :
Le **premier chapitre** représente l'état de l'art, dans lequel nous avons présenté les techniques d'analyse et d'interprétation de l'image qui constituent l'outil principal de la vision par ordinateur, on s'intéresse beaucoup plus sur la vision par ordinateur et ses niveaux (bas niveau et haut niveau).Ensuite, nous présentons les systèmes d'aide au diagnostic et ses étapes.
Le **deuxième chapitre** étudié le domaine d'expertise qui nous intéresse qui est le cancer du sein; en présentant les outils d'imagerie médicale, les pathologies mammaires et ses caractéristiques ainsi que les systèmes utilisés par les radiologues pour le dépistage du cancer du sein.
Dans le **troisième chapitre** nous présentons l'état de l'art des algorithmes génétiques, cette technique que nous l'avons utilisée pour le développement de notre approche de sélection de caractéristiques pertinentes des masses mammographiques.
Le **quatrième chapitre** présente une description détaillée de l'approche proposée, ainsi les résultats expérimentaux obtenus en appliquant notre méthode sur la base de caractéristiques calculés de l'image mammographique.
Enfin, ce mémoire est terminé par une conclusion où nous présentons nos perspectives.

Chapitre 1

Vision par ordinateur & Diagnostic assisté par ordinateur

1.1 Introduction

La vision artificielle est un processus de traitement de l'information. Elle utilise des stratégies bien définies afin d'atteindre ses buts. L'entrée d'un système de vision est constituée par une séquence d'images. Ce système apporte un certain nombre de connaissances qui interviennent à tous les niveaux. La sortie est une description de l'entrée en termes d'objets et de relations entre ces objets.

La vision par Ordinateur est à la base de tout système de vision artificielle. Un système de vision traite les informations de bas niveau d'abstraction de l'image pour en extraire des informations de plus haut niveau d'abstraction : présence d'entités, classe d'entités (document, chaise, mur) et plus précisément l'identité de l'entité.

Dans ce chapitre, on se focalise sur les phases de la vision par ordinateur qui sont divisées en problèmes de bas niveau et haut niveau ensuite, nous détaillons le processus de la reconnaissance de formes, et à la fin non se concentre sur le système d'aide au diagnostic et ses composants.

1.2 La Vision par ordinateur

La vision par ordinateur (aussi appelée vision artificielle) constitue une des branches de l'intelligence artificielle, c'est un domaine d'applications née à la fin des années 50 dont les premières bases théoriques ont été définies dans les années 60. Depuis, étant donné le spectre très large d'applications, très peu de problèmes ont trouvé des solutions entièrement satisfaisantes. La recherche en vision est divisée en plusieurs approches, chacune étudie un problème liée aux types d'images en utilisant des techniques différentes.

La vision par ordinateur 2D met l'accent sur les techniques actuelles d'analyse d'image, d'apprentissage et de reconnaissance des formes. La vision par ordinateur 3D approfondit quant à elle les approches permettant d'extraire d'une ou plusieurs images des informations

tridimensionnelles relatives à la scène photographiée. Par exemples l'industrielle, militaires, aérospatiales, la télédétection, la sécurité et le domaine médical [8].

La vision par ordinateur est la discipline qui cherche à reproduire la perception visuelle humaine sur un ordinateur. Le processus d'acquisition d'images est effectué à l'aide des caméras. Ces images seront traitées pour faire le passage de scène 2D à la scène 3D. Le but fondamental de la vision par ordinateur est de modéliser, reproduire, et surtout dépasser la vision humaine à l'aide de logiciels et de matériel à différents niveaux, Il a besoin de connaissances en informatique, en génie électrique, en mathématiques, en physiologie, en biologie et en sciences cognitives.

1.2.1 Difficulté de la vision par ordinateur

Parmi les difficultés les plus rencontrées de la vision artificielle :

- La perte d'informations suite au passage du 3D à 2D, en effet, c'est un phénomène qui se produit durant la capture d'image par un appareil photo.

-Le bruit est présent de façon inhérente dans chaque mesure au sein du monde réel. Son existence appelle aux outils mathématiques pouvant l'éliminer. En contrepartie, des outils plus complexes rendent l'analyse d'image beaucoup plus compliquée par rapport aux méthodes standards.

- Les quantités des images et des séquences vidéo sont énormes. Par exemple une feuille de papier A4 numérisée monochromatique à 300 points par pouce (ppp) à 8 bits par pixel correspond à 8,5 MB. Citons un autre exemple d'une vidéo RGB de 24 bits couleur 512x768 pixels, 25 images par seconde, rend un flux de données de 225 MB par seconde. Cela nécessite la conception d'un traitement simple et pertinent, dans le cas contraire, il serait difficile d'obtenir des performances en temps réel, c'est à dire, pour traiter 25 ou 30 images par secondes.

- La mesure d'intensité lumineuse ou du rayonnement dépend de l'éclairement énergétique c'est-à-dire le type de la source de lumière, la position de l'observateur etc. Un lien direct entre l'apparence des objets dans les scènes et leur interprétation est distingué.

-L'utilisation d'une fenêtre locale par l'ordinateur qui voit l'image à travers un trou de serrure, rend difficile la compréhension du contexte plus global, or que nous avons besoin d'une vision globale pour l'interprétation d'une image, en effet, les algorithmes d'analyse d'images analysent un bac de stockage particulier dans une mémoire opérationnelle et son voisinage

local; il est souvent très difficile d'interpréter une image si elle est perçue uniquement localement ou si seulement quelques trous de serrure locaux sont disponibles [1].

1.3 Analyse d'image

Le codage ou la représentation informatique d'une image implique sa numérisation. Cette numérisation se fait dans deux espaces :

- l'espace spatial où l'image est numérisée suivant l'axe des abscisses et des ordonnées : on parle d'échantillonnage. Les échantillons dans cet espace sont nommés pixels et leur nombre va constituer la définition de l'image.
- l'espace des couleurs où les différentes valeurs de luminosité que peut prendre un pixel sont numérisées pour représenter sa couleur et son intensité ; on parle de quantification. La précision dans cet espace dépend du nombre de bits sur lesquels on code la luminosité et est appelée profondeur de l'image.

La qualité d'une image matricielle est déterminée par le nombre total de pixels ("picturelement") et la quantité d'information contenue dans chaque pixel (souvent appelée profondeur de numérisation des couleurs).

1.3.1 Analyse de bas niveau

Les techniques de bas niveau de la vision par ordinateur représentent la base du traitement numérique de l'image [1], ils utilisent très peu de connaissances sur le contenu des images. Tout d'abord une image d'entrée est capturée par une caméra de télévision en (2D) et numérisée, étant décrite par une fonction d'image f(x, y) dont la valeur est en général l'intensité lumineuse en fonction de deux paramètres x, y, représentant les coordonnées de l'emplacement dans l'image. Puis l'étape du traitement qui consiste en les opérations effectuées sur les images au plus bas niveau d'abstraction comme l'élimination du bruit au sein de l'image, l'amélioration de certains descripteurs des objets jugés pertinents pour interpréter l'image. L'entrée et la sortie sont des images d'intensité. Ces images iconiques sont généralement de la même nature que les données originales capturées par le capteur, le traitement ne fait pas augmenter les informations du contenu de l'image, bien au contraire, il diminue généralement les informations contenues au sein de l'image. Du point de vue de la théorie de l'information, le meilleur traitement est de ne pas faire un traitement. La meilleure façon d'éviter l'élaboration d'un traitement est de se focaliser sur une acquisition d'images de haute qualité. Toutefois, le traitement est très utile dans une variété de situations, car il permet

de supprimer les informations qui ne sont pas pertinentes pour l'analyse de l'image. Par conséquent, le but du traitement est l'amélioration des données d'image en supprimant les indésirables distorsions et en améliorant certaines caractéristiques importantes de l'image pour l'analyse ultérieur de l'image, bien que les transformations géométriques d'images comme la rotation et le redimensionnement sont également classés comme des méthodes de traitement. Sachant que les pixels voisins correspondant à un objet donné dans des images réelles ont essentiellement la même valeur d'intensité lumineuse, de sorte que si un pixel déformé peut être capté à partir de l'image, il peut généralement être restauré comme une valeur moyenne des pixels voisins, ceci est un exemple d'un traitement que nous pourrions effectuer sur une image. La segmentation d'image est la prochaine étape du processus d'analyse d'image, dans laquelle l'ordinateur tente de séparer les objets de l'arrière-plan de l'image. Nous distinguons une segmentation totale et une segmentation partielle : la segmentation totale n'est possible que pour des tâches très simples, un exemple étant la reconnaissance d'objets non jointifs sombres du fond clair. Dans des cas de problèmes plus compliqués, les techniques d'analyse d'image de bas niveau gèrent les tâches de segmentation partielle, dans laquelle seuls les indices qui aideront l'analyse ultérieure de haut niveau sont extraits. La description d'objet dans une image totalement segmentée est également comprise dans le cadre d'analyse de bas niveau d'image.

1.3.2 Analyse de haut niveau d'image

L'analyse de haut niveau s'appuie sur des connaissances relatives au contenu de l'image[1], par exemple, taille de l'objet, sa forme et les relations mutuelles entre les objets dans l'image ; ces données de haut niveau sont généralement exprimées sous une forme symbolique. Les méthodes d'intelligence artificielle sont largement applicables aussi, en effet, la vision par ordinateur de haut niveau tente d'imiter la cognition humaine et la capacité à prendre des décisions en fonction de l'information contenue dans l'image. La vision de haut niveau commence par une certaine forme du modèle formel du monde, puis la réalité perçue sous la forme d'images numériques est comparée au modèle; l'ordinateur passe en analyse d'image à bas niveau pour trouver des informations nécessaires afin de mettre à jour le modèle. Ce processus est répété de manière itérative, et l'interprétation d'une image devient ainsi une coopération entre les processus top-down et bottom-up. Une boucle de retour d'information est introduite dans laquelle des résultats partiels de haut niveau créent des tâches pour l'analyse de bas niveau, et le processus itératif d'interprétation d'image devrait finalement converger vers l'objectif général.

1.4 La reconnaissance de forme

La reconnaissance de formes est une discipline dont le but est la classification des objets dans un grand nombre des catégories ou classes ; elle est également une partie intégrante dans le système d'intelligence artificielle construit pour prendre des décisions [2].

La reconnaissance des formes fait partie de la réduction méthodique d'information à partir d'une donnée très riche, par exemple : à partir d'une image numérisée, nous pouvons extraire une information pertinente qui tient en quelques bits, ou bien l'indication que l'image contient une forme circulaire ou rectangulaire. Nous considérons donc souvent la reconnaissance de formes comme un problème de classification, c'est-à-dire un problème de synthèse d'une fonction qui affecte chaque donnée prévisible à la catégorie pertinente.

Il peut s'agir de contenu visuel (code barre, visage, empreinte digitale) ou sonore (reconnaissance de parole), d'images médicales (échographie, rayon X ou IRM...) multi spectrales (images satellitaires) ou bien d'autres. La RDF s'intéresse à la conception et à la réalisation de systèmes (matériels et logiciels) capables de percevoir, et dans une certaine mesure, d'interpréter des signaux captés dans le monde physique [3].

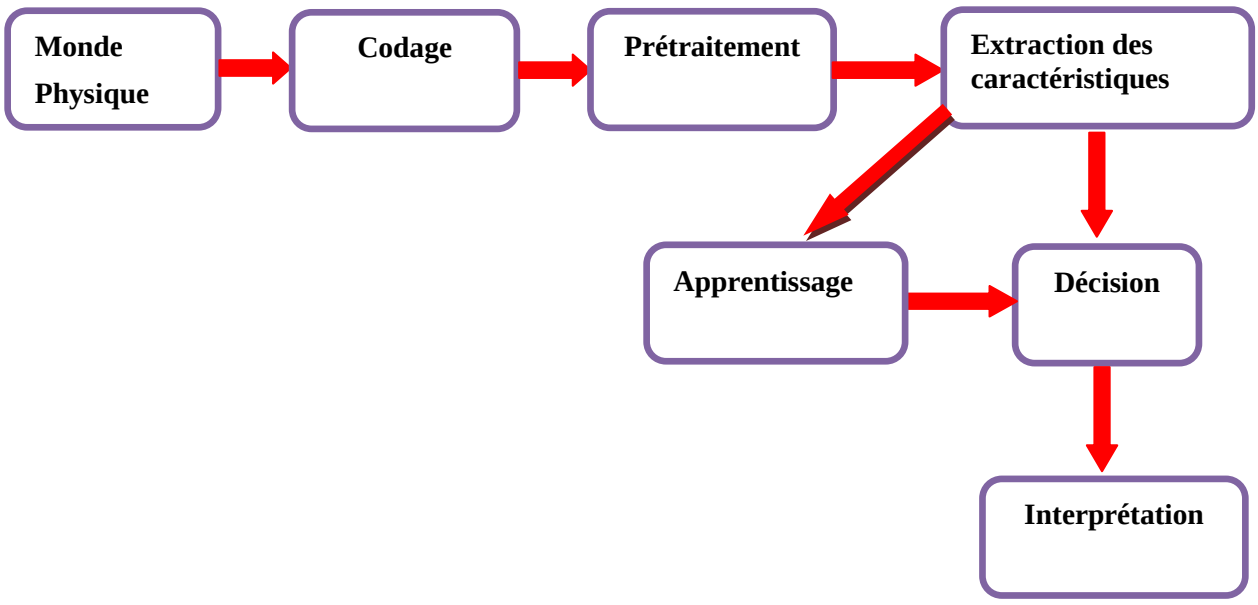

Figure 1.1 : *Processus de la RDF.*

1.4.1 Principe des RDF

La RDF est un des nombreux aspects de l'intelligence artificielle. A partir d'un ensemble de données ou d'informations apprises, elle offre la possibilité d'interpréter toute nouvelle

observation(ou forme). Les observations déjà connues sont regroupées en classes, constituant des prototypes auxquels la nouvelle observation est comparée pour être identifiée. Les algorithmes utilisés permettent donc de classer des observations dont les propriétés ont varié par rapport à une observation type [5].

1.4.2 Processus de RDF

Classiquement, un processus de Reconnaissance des Formes (RDF) extrait des Interprétations à partir des Représentations d'une image. La qualité d'un processus de RDF admet deux caractéristiques induites par le lien entre Représentation et Interprétation : sa complexité et sa complication. Ainsi, dans le cas d'objets manufacturés, les modèles sont clairement établis et la solution au problème de reconnaissance, pour compliquée qu'elle soit, n'en est pas moins déterminée [5].Le processus de RDF est un processus de réduction progressive et sélective de l'information. Les étapes de traitement d'un processus de RDF (Voir Figure 1.1).

1.4.2.1 Monde physique

Le monde physique qui nous entoure est considéré comme un espace analogique de dimension *n* appelé l'espace de formes *F*. C'est celui qui est présenté dans sa forme la plus primaire, c'est-à-dire dont nous devons déterminer les caractéristiques les plus apparentes avant l'étape du codage [6].

1.4.2.2 Codage

Est une opération qui consiste en une conversion numérique du monde physique continu vers un monde discret.

1.4.2.3 Prétraitement

Elle permet de sélectionner, dans l'espace de représentation, l'information nécessaire à l'application. Cette sélection passe par l'élimination du bruit dû aux conditions d'acquisition, par la normalisation des données ainsi par la suppression de redondance.

1.4.2.4 Extraction des caractéristiques

L'objectif de l'extraction et de la sélection de caractéristiques est d'identifier les caractéristiques importantes pour la discrimination entre classes. Après avoir choisi le meilleur ensemble de caractéristiques, il s'agit de réduire la dimensionnalité de l'ensemble des

caractéristiques en trouvant un nouvel ensemble, plus petit que l'ensemble original, qui néanmoins, contient la plupart de l'information.

Quelques exemples sont présentés :

-En reconnaissance des caractères, les caractéristiques utilisables peuvent venir de la densité des points, des moments, des lieux caractéristiques, des transformées mathématiques (Fourier, Walsh, Hadamard...), elles peuvent également venir des squelettes ou des contours.

-Dans des applications liées à l'analyse de texture telles que télédétection et analyse des scènes, les caractéristiques utilisables peuvent venir de la matrice déco-occurrence, des descripteurs de Fourier, du spectre de puissance, des moments, aussi bien que de diverses primitives structurelles.

-Dans l'analyse et la reconnaissance de formes d'ondes telles que le signal sismique, l'EEG et l'ECG, la parole aussi bien que les images de formes courbes, les caractéristiques utilisables peuvent venir du spectre de puissance, de fonctions d'approximation, des zero crossing, et de plusieurs types de segments de traits structurels [17].

1.4.2.5 Apprentissage

L'apprentissage tente de définir des classes de décision ou d'appartenance. Son rôle est déclarer la décision à l'aide de connaissance à priori sur les formes, à partir de critères spécifiques aux formes. Il existe 2 types d'apprentissages (supervisé, non supervisé) [4].

a)**Apprentissage supervisé :** Les exemples d'apprentissage sont étiquetés afin d'identifier la classe à laquelle ils appartiennent. Le but de l'algorithme de classification est de classifier correctement les nouveaux exemples dans les classes définies dans la phase d'apprentissage. Les méthodes d'apprentissage supervisé se construisent à partir de la base d'apprentissage, des classifier, ou fonctions de classement [7].

Si l'utilisateur possède suffisamment d'informations sur la population à étudier, il peut effectuer une classification supervisée. Cette catégorie suppose avoir un groupe d'individus de chaque classe, dont on connaît leur appartenance. Ces individus forment des échantillons "d'apprentissage". Ils sont utilisés pour entraîner le classifier. D'autres échantillons, dits "de test", servent à valider la classification en évaluant sa pertinence à travers le taux d'individus bien classés. Il existe plusieurs méthodes de classification supervisées.

Les méthodes les plus réputées sont l'analyse discriminante linéaire, la régression logistique, les réseaux de neurones. Les réseaux de neurones artificiels, connus par l'acronyme anglais

ANN (Artificial Neural Networks), sont largement utilisés pour les problèmes de classification. Ils reposent sur la théorie des perceptrons. Un ANN est composée de plusieurs neurones répartis sur une couche d'entrée (désignant les descripteurs), une couche de sortie (désignant le résultat de classification) et un nombre de couches cachées. Toutefois, l'inconvénient de cette méthode est le choix du nombre de couches cachées et du nombre de neurones dans chaque couche. Ainsi, l'utilisateur est amené à faire des essais avec différentes combinaisons du nombre de couches et de neurones afin d'aboutir au réseau de neurones le plus adapté à son type d'application.

b) **Apprentissage non supervisé :** L'algorithme d'apprentissage cherche à trouver des régularités dans une collection d'exemples, puisque dans ce type d'apprentissage on ne connaît pas la classe à laquelle les exemples d'apprentissage appartiennent. Une technique employée consiste à implémenter des algorithmes pour rapprocher les exemples les plus similaires et éloigner ceux qui ont le moins de caractéristiques communes [4].

Ces techniques sont utilisées lorsque l'identité des classes n'est pas connue. Cela résulte d'un manque d'information de la population à étudier. La classification non-supervisée, dite automatique, ou groupement connue par (clustering en anglais) consiste à déterminer les différentes classes naturellement sans aucune connaissance préalable.

Parmi les méthodes de classification non supervisées la méthode la plus communément utilisée est celle de l'algorithme K-moyennes également appelée algorithme des nuées dynamiques (en anglais k-means) [9].L'algorithme fonctionne en précisant le nombre K de classes (clusters) attendues (K étant fixé par l'utilisateur). Il calcule la distance intra-classe et refixe les centres de classe selon les valeurs de distance. Les inconvénients de cette méthode sont premièrement la nécessité de fixer le nombre de classes avant de commencer la classification. Deuxièmement, cette méthode est très sensible à la répartition initiale des données.

1.4.2.6 Décision :

La décision ou classement est l'étape proprement dite la reconnaissance son rôle est de classifier la forme ciblée à partir de l'apprentissage réalisée. Pour la décision et pour l'apprentissage, les critères utilisés sont habituellement les mêmes.

1.4.3 Domaines d'application

La reconnaissance de formes est utilisée dans plusieurs domaines d'activités. Parmi ces domaines on peut citer :

- Recherche d'images par le contenu : Cette technologie est actuellement intéressante pour la recherche de données sur l'imagerie médicale, ou cartographiques.
- Classification de documents : Son activité est essentielle dans de nombreux domaines économiques : elle permet d'organiser des corpus documentaires, de les trier, et d'aider à les exploiter dans des secteurs tels que l'administration, l'aéronautique, la recherche sur internet, les sciences.
- Reconnaissance de l'écriture manuscrite : Cette technologie fait appel à la reconnaissance de forme, mais également au traitement automatique du langage naturel. Cela veut dire que le système, tout comme le cerveau humain, reconnaît des mots et des phrases existant dans un langage connu plutôt qu'une succession de caractères. Ceci améliore grandement la robustesse.

1.5 Diagnostic Assistée par Ordinateur (DAOx)

Diagnostic assisté par ordinateur (DAOx), sont des systèmes qui aident les médecins dans l'interprétation des images médicales. Les techniques d'imagerie en radiologie, en IRM et en échographie donnent beaucoup d'informations que le radiologue ou un autre professionnel doit analyser et évaluer de façon exhaustive en peu de temps. Les systèmes de DAO traitent des images numériques pour des apparitions typiques et mettent en évidence des sections visibles, telles que des maladies possibles, afin d'offrir des apports à l'appui d'une décision prise par le professionnel.

1.5.1 Composition des systèmes DAOx

En pratique, le système de diagnostic assistée par ordinateur (DAOx) dédié à l'analyse d'images mammographies est une suite de phases qui doivent être exécutées l'une après l'autre, depuis l'acquisition de l'image jusqu' à la prise de décision. Certaines de ces phases sont souvent étroitement liées et indissociables. Les étapes de traitement d'une image mammographie peuvent se résumer en :

- une étape de prétraitement qui sert à améliorer la qualité de l'image avant toutes manipulations.
- une étape de segmentation qui permet de détecter la lésion à étudier.
- une étape de description qui a pour but de caractériser les lésions à travers des formulations mathématiques.
- une étape de classification et de prise de décision en utilisant un classeur adéquat.

Ces différentes étapes sont résumées dans le diagramme représente dans la figure 1.2.

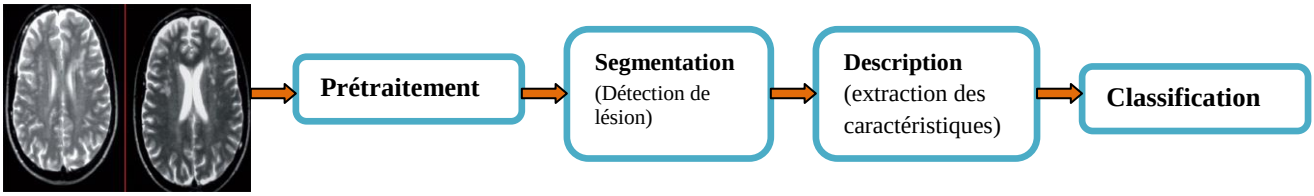

Figure 1.2 : *Diagramme général d'un système DAOx.*

1.5.2.1 Prétraitement

Le cancer du sein (comme tous les cancers d'une manière générale) doit être détecte dans sa première phase pour maximiser les chances de survie. Sauf que dans cette phase, il s'avère très difficile de repérer, à l'œil nu, la pathologie dans le tissu mammaire environnant sans avoir recours à un prétraitement spécifique de l'image acquise. D'où l'objectif principal de cette étape est d'augmenter le contraste entre la lésion mammaire (que ce soit masse ou micro-calcification) et le reste de l'image pour faciliter les traitements ultérieurs. Sachant que dans le cas où une région d'intérêt diffère en luminance de moins de 2% du reste de l'image, elle demeure indiscernable à l'œil nu [11]. Le prétraitement des images mammographies est connu sous le nom de rehaussement ou d'amélioration du contraste. Le problème majeur des algorithmes de rehaussement du contraste réside dans le fait que certaines régions peuvent ne pas être rehaussées convenablement alors que d'autres peuvent être l'objet d'un excès de rehaussement. Un manque de rehaussement du contraste peut causer des faux négatifs (FN). En effet, plusieurs détails concernant la lésion peuvent être négligés. Dans ce cas, certaines lésions peuvent ne pas être détectées et par la suite non diagnostiquées. Ce qui ne répond pas à l'objectif principal de la détection précoce d'un cancer. Un excès de renforcement du contraste peut causer des faux positifs (FP). Dans ce cas, plusieurs détails inexistants réellement peuvent s'ajouter à la lésion. D'où certaines régions normales du tissu mammaire peuvent être considérées comme lésions ce qui va engendrer des biopsies inutiles. Les techniques traditionnelles d'amélioration du contraste ont été appliquées à la mammographie pendant plus de trois décennies.

Une approche couramment utilisée pour l'amélioration du contraste est la modification globale de l'histogramme. Cette méthode consiste à réaffecter les valeurs d'intensité des pixels afin de rendre la nouvelle répartition des intensités plus uniforme. Ceci peut êtreréalise par l'égalisation d'histogramme ou par l'étirement d'histogramme [12]. Toutefois, ces transformations ont l'inconvénient de faire ressortir le bruit dans les images déjà bruitées.

1.5.2.2 Segmentation

Il n'existe pas de définition précise (mathématique) et unique de la segmentation. La définition de la segmentation fait appel à notre sens commun de la perception des choses. En somme, segmenter revient à diviser l'image en zones d'intérêt afin de réduire sa complexité.

La segmentation est souvent considérée e comme l'étape initiale dans un système de diagnostic assiste par ordinateur (DAOx) surtout si on fait abstraction de l'étape e de prétraitement qui, d'après la section précédente, n'est pas indispensable dans le cas de traitement des masses. La phase de segmentation est très importante puis que les traitements ultérieurs (description et classification) sont fortement lies au résultat de segmentation. En effet, une bonne détection du contour de la lésion engendre une description fidèle à ses caractéristiques. Ainsi, on peut garantir une classification minimisant le taux des faux positifs et maximisant le taux des vrais négatifs.

Cette étape consiste à détecter ou bien la totalité du sein à partir du fond de l'image ou bien un type d'anomalie bien spécifique comme les micros calcifications et les masses. Il aété démontré que la détection des mas ses est plus difficile que la détection des microcalcifications. En effet, les masses peuvent être masquées partiellement par le tissu mammaire.

Il existe deux grandes approches de segmentations à savoir l'approche contour et l'approche régions.

- ✓ l'approche contour est basée sur la détection des frontières qui forment la région d'une image .les méthodes peuvent soit modéliser le contour et essayer de rechercher dans l'image les pixels qui répondant à ce modèle, soit rechercher les points localisant une rupture de modèle de régions adjacents.
- ✓ L'approche région consiste à déterminer directement les régions en regroupant les points qui ont les mêmes propriétés statiques ou structurelles.

a) Méthodes de détection contours

Les méthodes de segmentation par détection de contours recherchent à localiser les transitions entre les régions. Elles utilisent la discontinuité dans une image pour détecter les bords et les contours régions. Les contours se manifestant dans l'image par une forte transition des valeurs de l'intensité (fort contraste).

Les méthodes dérivatives utilisées se basent sur des opérateurs tels que Roberts, Sobel, Prewitt et Canny. Une fois la norme et la direction du gradient sont calculées en chaque pixel de l'image, ces méthodes extraient des contours d'un seul pixel d'épaisseur en sélectionnant les maximas locaux des normes des gradients. En mammographie, on est souvent confronté à la présence de bruit (d'acquisition), de textures fines ou de frontières pas très nettes, d'où les transitions détectées ne correspondent pas forcément aux contours réels. Il est alors souvent nécessaire d'appliquer un traitement en aval an d'écarter les transitions dues aux bruits. De plus, les contours extraits sont généralement discontinus et peu précis. Il faut donc, utiliser des techniques de reconstruction de contours par interpolation ou connaître a priori la forme de l'objet recherché an de connecter les points du contour.

Dans le cas des approches région, plusieurs travaux [15] se basent sur le fait que les intensités sont statistiquement homogènes dans chaque région à segmenter. Or l'inhomogénéité de l'intensité se produit souvent dans les images médicales comme c'est le cas des images mammographies. Le modèle du contour actif basé région récemment proposé par [16] est capable de segmenter des images ayant diverses intensités inhomogènes. Par ailleurs, il parvient à fournir un bon résultat de segmentation dans le cas d'objets à contours mal dénis ou masqués (ce qui est souvent le cas des masses mammaires). En utilisant le terme de régularisation proposé par [16], la régularité de la fonction "level set" est intrinsèquement préservée ce qui garantit la précision du calcul et évite les procédures coûteuses de réinitialisation.

b) Méthodes de détection région

Les méthodes de l'approche région cherchent à différencier les régions en utilisant les propriétés de l'image telle que la couleur, texture, forme … etc. ces méthodes utilisant principalement les critères de décision pour segmenter l'image en différents régions selon la similarité des pixels. Nous proposons dans la suite une méthode de segmentation de type région.

Méthode de segmentation par seuillage: Le seuillage a pour objectif de segmenter une image en deux ou plusieurs classes. Cette opération consiste à effectuer une partition de l'histogramme en niveaux de gris en utilisant un ou plusieurs seuils. Chaque pic de l'histogramme correspond à une classe. En effet, cette méthode n'est efficace que si l'histogramme contient réellement des pics séparés. Les méthodes de seuillage ont été largement utilisées pour la segmentation de masses mammaires. Par exemple, les auteurs dans

[13] ont utilisé différentes valeurs de seuils en niveau de gris qui dépendent du type de tissu mammaire et ceci en se basant sur une analyse de l'histogramme.

Plus récemment, MUDIGONDA, Naga R., RANGAYYAN, Rangaraj M., et DESAUTELS, JE Leo [14] ont utilisé un seuillage multiniveaux pour détecter des contours fermés. L'inconvénient majeur de cette approche est le fait de considérer que les masses ont une densité uniforme par rapport au fond de l'image ce qui n'est pas toujours vérifié.

Cependant, l'inconvénient majeur de ces différentes techniques de seuillage est le choix du seuil ou de l'intervalle de seuillage. En effet, avec un intervalle trop large, on obtient des faux positifs. Dans ce cas, l'image seuillée contient des pixels qui ne font pas partie des objets d'intérêt. Il s'agit généralement de bruit ou de pixels (zones) qui ont un niveau de gris proche de celui des objets recherchés. Avec un intervalle trop étroit, on obtient des faux négatifs. Certains objets d'intérêt ou bien des parties de ces objets n'apparaissent pas dans l'image seuillée.

1.5.2.3 Description

L'être humain reçoit en permanence des informations très diverses et très complexes par l'intermédiaire de ses cinq sens. En dépit de l'abondance de ces informations, le cerveau humain est capable de restituer chaque objet observé et de lui attribuer une représentation cohérente appelée "description humaine".

Dans le domaine du traitement de l'image, la description est l'étape qui cherche à reproduire le même processus d'analyse et d'interprétation. En effet, la description a pour but d'extraire les caractéristiques qui décrivent au mieux et de façon quantitative ou qualitative les objets présents dans l'image. Elle transforme les informations de bas niveau issues de la phase d'acquisition (après probablement prétraitement et segmentation) en informations de haut niveau de telle sorte que les formes et les structures soient décrites de façon analytique. De manière générale, plus la description effectuée est proche de "la description humaine", plus elle est considérée comme robuste et fidèle à l'image initiale.

Les méthodes de description d'images sont variables et dépendent de l'objectif visé (description globale, description locale) et du type d'image à analyser (image binaire, image en niveau de gris, image couleur). En littérature, la description d'images est assurée en utilisant la couleur, la texture et/ou la forme.

1.5.2.4 Classification

La classification est considérée comme la dernière étape dans un système d'aide au diagnostic. Elle exploite le résultat de description pour pouvoir décider de la nature pathologique de la masse.

La notion de classification signifie l'affectation d'une étiquette à des échantillons d'une base de données en utilisant un certain nombre de caractéristiques. Ces caractéristiques doivent bien évidemment être capables d'identifier chaque échantillon. En traitement d'images, l'échantillon peut désigner un pixel, une zone dans l'image, un objet représenté dans l'image ou l'image elle-même. Selon l'application, le but de la classification est soit de :

-Classifier les pixels de l'image en différentes zones. Dans ce cas, le problème de classification revient à un problème de segmentation d'images en différents objets.

A titre d'exemple, on peut classifier les différentes zones d'une image mammographie en lésion ou non lésion.

-Classifier l'image ou les objets de l'image selon différentes catégories. A titre d'exemple, on peut classifier les masses qui se trouvent dans les images mammographie en malignes ou bénignes.

Nous distinguons deux catégories de méthodes de classification: les classifications non supervisées et celles supervisées qui sont présentées précédemment dans la section (1.4.2.5).

1.6 Conclusion

La vision par ordinateur est un outil de traitement de l'information, texte, image …etc. Dont le but de faire une relation entre la compréhension de l'image (analyse) et la manipulation ou modification par différentes techniques pour un seul objectif : Analyser l'information et faire des traitements pour améliorer l'information et interpréter l'information (image).

Chapitre 2

Classification des anomalies mammaires

2.1 Introduction

Afin de proposer des outils performants aidant les radiologues à dépister le cancer du sein, les chercheurs ont conçu des systèmes de détection et de classification automatiques des lésions mammaires. Nous présentons une étude de ces systèmes et nous concentrons sur la description et la classification automatiques des lésions dans les images mammographiques.

En effet, nous commençons, dans ce chapitre, par présenter les outils d'imagerie médicale permettant le dépistage et le diagnostic de ce type de cancer notamment la mammographie. Ensuite, nous étudions les pathologies mammaires et la classification de ces pathologies selon la norme BI-RADS(Breast Imaging-Reporting And Data System).

2.2 Imagerie mammaire

L'imagerie a une place fondamentale dans la prise en charge des lésions mammaires dans toutes les étapes du dépistage jusqu'au suivi après traitement. Il existe actuellement, un certain nombre de techniques d'imagerie du corps humain couramment employées dans le domaine médical. Chacune d'elles est sensible à un type de contraste particulier et trouve ses applications pour des organes différents. Plusieurs techniques peuvent, également apporter des informations complémentaires sur un même organe. Les outils de l'imagerie médicale utilisés pour le dépistage ainsi que le diagnostic du cancer du sein sont : l'échographie (imagerie par ultrasons), l'IRM (Imagerie par Résonance Magnétique) et la mammographie (imagerie par rayons X). En effet, on explique ces différentes techniques ainsi que leurs caractéristiques.

2.2.1 La mammographie

La mammographie est la méthode moderne utilisée pour dépister le cancer du sein .une mammographie est un radiographie des seins qui est un objet tridimensionnel, et de référence des lésions du sein, Il permet d'obtenir des images de l'intérieur du sein à l'aide de rayons X et de détecter ainsi certaines anomalies. Une mammographie est pratiquée dans deux circonstances : d'un dépistage ou d'un diagnostic précoce du cancer du sein. Quelles que

soient les circonstances, deux clichés (photos) par sein sont réalisés, un cliché de face et un en oblique, ce qui permet de comparer les deux côtés de chaque sein. Dans l'immense majorité des cas, est le premier examen d'imagerie.

Les indications de la mammographie sont surveillance des seins déjà traités pour un cancer, un bilan ou une surveillance d'un traitement hormonal substitutif, un écoulement mamelonnaire. Elle est également prescrite dans le cadre du dépistage individuel du cancer du sein. La mammographie permet de mettre en évidence des anomalies (opacités, calcifications) qui, en fonction de certains paramètres nombre, évolution, etc.), orientent vers une pathologie bénigne ou vers un cancer.

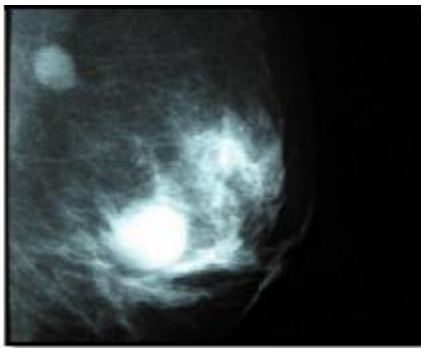

***Figure 2.1** : Lésion d'une mammographie [27].*

L'appareil dédié à la réalisation d'une mammographie est le mammographe (figure 2.2). Cet appareil se compose d'un tube radio gène générateur de rayons X de faible énergie et d'un système de compression du sein. En premier temps, les deux seins sont comprimés à tour de rôle. Cette compression permet l'étalement des tissus mammaires ce qui facilite la visualisation des structures du sein et la réduction de la dose de rayons X délivrée. En deuxième temps, les deux seins sont e exposés à une faible dose de rayons X. On obtient, alors, une projection du sein sur un détecteur plan. La radiographie est réalisée sur des films argentiques ou sur des systèmes s de radiologie digitale de haute qualité. L'analyse de la glande mammaire est réalisée grâce aux différences de l'atténuation des différents types de tissu. Nous détaillons dans la section suivante l'anatomie du sein ce qui permet par la suite d'établir la relation entre la nature du tissu mammaire et l'infiltration des rayons X.

2.2.2 Échographie

L'échographie est une technique d'imagerie médicale qui repose sur l'utilisation d'ultrasons qui ne délivre pas de rayons X, et permet de « visualiser » l'intérieur du corps. Cet examen produit des images en temps réel (c'est-à-dire visibles immédiatement) de l'intérieur du sein.

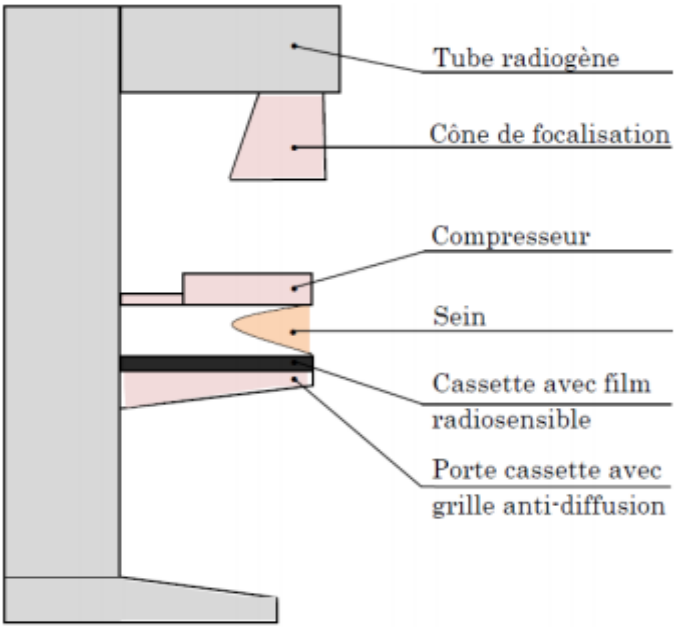

***Figure 2.2** : Les composants d'une mammographie.*

Cette technique d'imagerie s'applique sur les seins en utilisant un gel (hypoallergénique) permettant un bon contact entre la peau et la sonde d'échographie. Un examen échographique dure en moyenne 5 à 10 minutes. Cette technique est très utile pour voir la nature (liquide ou solide) des nodules palpés ou découverts sur une mammographie, Cet examen complète et précise les images obtenues par mammographie. Il ne remplace pas une mammographie qui est l'examen de référence pour la détection du cancer du sein. Pour les adolescentes, les jeunes femmes et les femmes enceintes, la mammographie est moins performante : l'échographie mammaire est donc l'examen de référence. Pour les femmes plus âgées, la mammographie reste l'examen de référence. L'échographie permet ainsi de comparer ce que l'on sent avec les doigts (lors de la palpation) et ce que l'on voit sur l'écran[28].

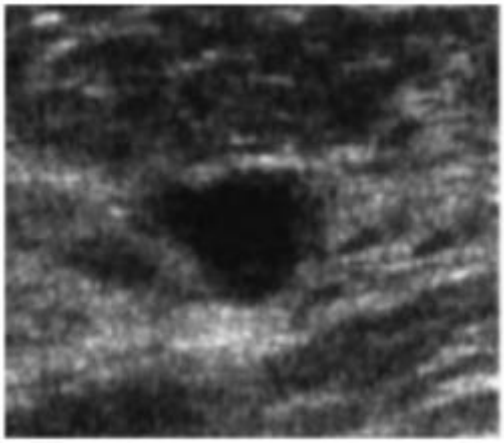

***Figure 2.3** : Image échographique du sein [5].*

La technique d'échographie présente deux avantages majeurs. Elle est d'une part peu coûteuse et d'autre part non invasive. Elle ne présente aucun risque pour la patiente, pour cette raison, elle est généralement utilisée dans le cas où la patiente est enceinte. Toutefois, vu certains inconvénients, l'échographie du sein n'est pas systématique. Elle est utilisée en complément d'une mammographie surtout qu'elle ne révèle que très rarement un cancer non détecté par la mammographie.

En outre, les microcalcifications qui sont de petites tailles (et même les petites masses) sont difficiles à détecter par ultrasons. Il est ainsi difficile de s'assurer que le sein ait été diagnostiqué dans sa totalité à l'issue de ce type d'examen. De plus, les images ultrasonores sont généralement altérées par un bruit spécifique appelé la granularité (reconnu aussi sous le nom speckle) qui est causé essentiellement par les interférences entre les ondes. En conséquence, cette méthode d'imagerie médicale n'est pas généralisée aux campagnes de dépistage. Elle est souvent exploitée comme moyen de repérage lors d'une ponction ou d'une biopsie et aussi comme moyen de repérage préopératoire pour marquer l'emplacement de la lésion.

2.2.3 Imagerie par Résonance Magnétique (IRM)

L'Imagerie par Résonance Magnétique est une technique d'imagerie médicale relativement récente (début des années 1980). Cette méthode se base sur l'utilisation d'un aimant(Constituant le champ magnétique) et d'ondes de radiofréquences. Son principe consiste à faire vibrer de façon imperceptible les atomes d'hydrogène du corps humain. Placés dans un champ magnétique puissant, tous les atomes d'hydrogène s'orientent vers la même direction.

L'IRM mammaire doit être bilatérale avec utilisation d'antennes dédiées au sein, elle a pour Le but de détecter et de caractériser des lésions mammaires. Cet examen nécessite impérativement une injection intraveineuse de produit de contraste. Durant l'examen, la patiente est couchée sur le ventre. Donc cette technique est polyvalente et fiable dont les deux inconvénients majeurs sont la faible disponibilité des élevé des examens. Elle ne dispense pas des techniques classiques d'imagerie, mais elle est particulièrement utile en cas d'hésitation diagnostique sur la bénignité d'une anomalie et pour le diagnostic des lésions multifocales.

Elle jouera peut être un rôle dans le dépistage du cancer du sein chez les femmes à très haut risque familial. En raison de sa grande sensibilité, une lésion dont l'image n'est pas renforcée par l'injection de gadolinium est quasi certainement bénigne[18].

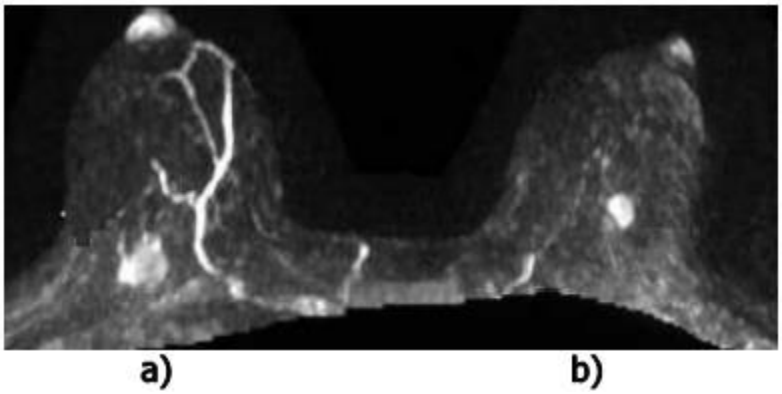

***Figure 2.4** : Image IRM transversal de la poitrine du patient[20].*

2.3 Les anomalies mammaires

Les cancers du sein apparaissent d'habitude avec des structures canalaires déformées. Il y a trois types majeurs de cancer du sein : les masses circonscrites/ovales, les lésions spiculées ou distorsions architecturales et les microcalcifications. Les lésions malignes ont généralement une forme plus irrégulière, ou mal définie, que les lésions bénignes.

2.3.1 Les micro calcifications

Une micro calcification est un dépôt de sels de calcium composé des substances chimiques. Ces substances sont très radio-opaques et se traduisent, dans les clichés mammographiques, par de petits points clairs. Les caractéristiques qui distinguent les microcalcifications des autres éléments sont leur fort contraste et leur petite taille (<0 ,5 *mm*). Une fois leur taille dépasse 1 *mm*, on les appelle des macrocalcifications et elles sont souvent bénignes. Les microcalcifications n'ont pas de taille minimale, ce qui fait que les plus petites d'entre elles peuvent facilement être confondues avec le bruit présent dans les images de mammographie. La description des microcalcifications permettant de décider de leur degré de suspicion inclut simultanément le critère de morphologie, de distribution et de nombre.

2.3.1.1 Morphologie

L'analyse de la morphologie est très importante **[21].** Elle permet le plus souvent de séparer les micros calcifications bénignes et malignes. Les microcalcifications arrondies ou ovales, uniformes dans leur taille et leur forme, sont souvent bénignes. A l'inverse, celles qui sont irrégulières et hétérogènes sont souvent malignes**[22]**.

On détaille dans ce qui suit les différents types de microcalcifications et on donne des exemples explicatifs de chaque cas dans la figure 2.5.

a) **Micro calcifications cutanées ou dermiques :** elles présentent typiquement un centre clair. Des clichés en incidence tangentielle sont souvent utilisés pour confirmer la localisation cutanée de ces microcalcifications.

b) **Micro calcifications vasculaires :** ces microcalcifications en rails ou linéaires sont associées à des structures tubulaires.

c) **Microcalcifications grossières ou coralliformes :** elles sont de grande taille (supérieures à 2-3 mm de diamètre).

d) **Microcalcifications en bâtonnets :** elles sont généralement associées à une ectasie canalaire (dilatation du canal galactophore) et sont alors dirigées vers le mamelon. Elles mesurent habituellement plus de 1 mm de large et peuvent présenter un centre clair si le dépôt calcique se fait dans la paroi du canal.

e) **Microcalcifications rondes** : elles ont une forme ronde et peuvent être de tailles variables. Lorsqu'elles mesurent moins de 0 .5 mm, elles sont dites punctiformes ou pulvérulentes.

f) **Microcalcifications à centres clairs** : leur taille peut s'étendre de 1 mm à plus de 1 cm. Elles sont rondes ou ovales, à surface lisse et à centre clair. La paroi calcifiée qui les entoure est plus épaisse que celle des microcalcifications en coquille d'œuf.

g) **Microcalcifications en coquille d'œuf ou pariétales :** ces microcalcifications très fines apparaissent comme des dépôts calciques sur la surface d'une sphère. Vu dans l'axe du rayonnement X, ce dépôt mesure généralement moins de 1 mm.

h) **Microcalcifications à type de lait calcique:** elles sont sédimentées dans le fond de kystes. En utilisant l'incidence cranio-caudale, elles sont souvent difficiles à discerner. Par contre, l'incidence de profil permet de démontrer leurs formes caractéristiques : semi-lunaires, en croissants, curvilignes ou linéaires.

i) **Microcalcifications de suture** : elles correspondent à des dépôts calciques sur du matériel de suture. Ces microcalcifications sont typiquement linéaires ou tubulaires et présentent fréquemment des nœuds.

j) **Microcalcifications dystrophiques** : elles mesurent habituellement plus de 0.5 mm de diamètre et sont de formes irrégulières. Elles présentent parfois un centre clair. Ces microcalcifications sont souvent rencontrées dans un sein irradié ou après un traumatisme mammaire. Elles représentent la majorité des cas retrouvés en pathologie mammaire.

k) **Microcalcifications amorphes ou indistinctes:** elles sont souvent plus ou moins rondes ou en forme de flocons. Elles sont de petites tailles et généralement à contours vagues sans forme spécifique.

l) **Microcalcifications fines et polymorphes:** elles sont habituellement mieux visibles que les microcalcifications amorphes. Elles sont irrégulières de taille et de forme variables mesurant généralement moins de 0.5 mm de diamètre.

m) **Microcalcifications linéaires et ramifiées:** elles mesurent moins de 0.5 mm d'épaisseur. Elles sont irrégulières et de formes parfois linéaires ou curvilignes généralement discontinues, coudées ou branchées.

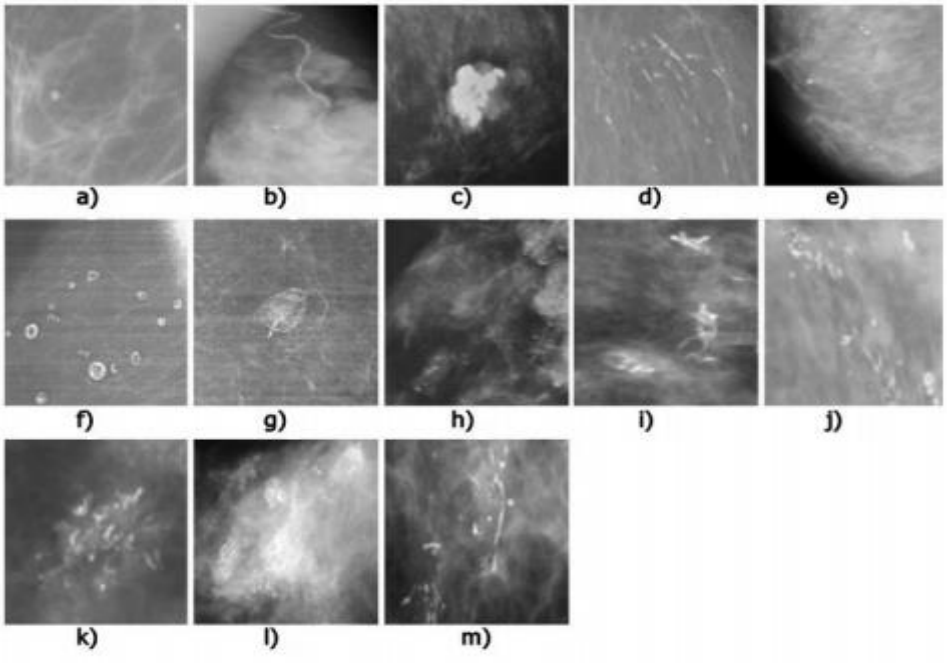

***Figure 2.5** : Les différents types de micro calcifications [19].*

2.3.1.2 Distribution

La distribution des microcalcifications est un critère fondamental. Elle présente leur répartition dans le sein et joue un rôle important dans la prise de décision de la malignité [22], Les différentes distributions possibles des microcalcifications sont détaillées dans ce qui suit et sont représentées dans la figure 2.6.

a) Microcalcifications diffuses/éparses : Dans ce cas, les microcalcifications sont distribuées de façon aléatoire dans l'ensemble du sein.

b) Distribution régionale : les microcalcifications sont dispersées dans un large volume du tissu mammaire (un ou plus d'un quadrant) et ne présentent pas une distribution canalaire.

c) **Microcalcifications groupées, en amas ou en cluster :** ces termes sont utilisés lorsque de multiples microcalcifications (au moins cinq) occupent un petit volume tissulaire.

d) **Distribution linéaire :** les microcalcifications sont disposées les unes derrière les autres sous forme d'une ligne. Il s'agit généralement de dépôts calciques dans un galactophore.

e) **Distribution segmentaire :** elle suggère des dépôts calciques dans des canaux galactophores ainsi que leurs branches ce qui évoque la possibilité d'un cancer mammaire étendu.

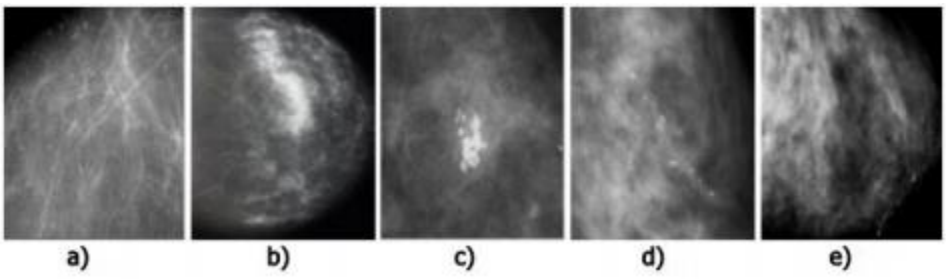

***Figure 2.6** : Les différentes distributions des microcalcifications[20].*

2.3.2 Les masses

Une masse (ou opacités) est une lésion importante occupant un espace et vue sur deux incidences différentes. Si une opacité potentielle est vue seulement sur une seule incidence alors elle est appelée asymétrie jusqu'à ce que son caractère tridimensionnel soit confirmé. Elles sont caractérisées par des descripteurs concernant leur forme, leurs contours, leur densité. Pour chaque descripteur.

2.3.2.1 La forme

Selon la description du BI RADS [19], On distingue généralement quatre classes de formes pour les opacités : les opacités rondes, ovales, lobulaires ou irrégulières. Ces différentes formes sont illustrées de manière schématique à la figure 2.7.

a) **Ronde** : Il s'agit de masse sphérique, circulaire ou globuleuse.

b) **Ovale** : Elle présente une forme elliptique (ou en forme d'œuf).

c) **Lobulée** : La forme de la masse présente une légère ondulation.

d) **Irrégulière** : Cette appellation est réservée aux masses dont la forme est aléatoire et ne peut être caractérisée par les termes cités ci-dessus.

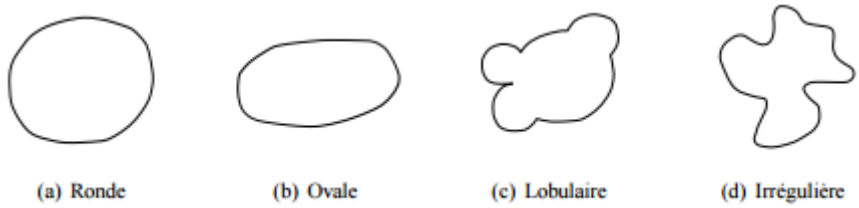

***Figure 2.7** : Différentes formes pour les opacités[29].*

2.3.2.2 Le contour

On dénombre cinq grandes classes de contours [23]. Est soit circonscrit, soit micro-lobulé, soit masqué soit indistinct, soit spiculé. On détaille dans ce qui suit ces différentes notions (figure 2.8).

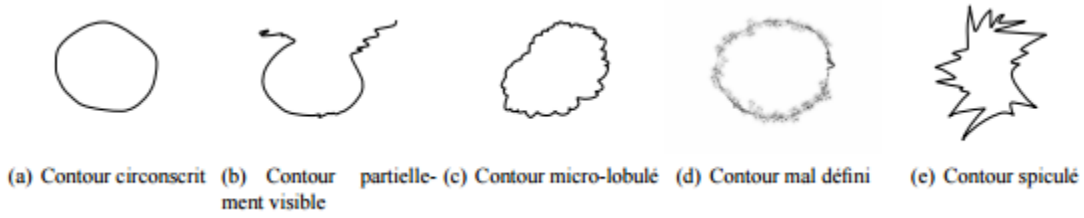

***Figure 2.8** : Différents types de contours pour les opacités [29].*

a) Circonscrit : qui sont des contours bien définis où la frontière entre lésion et fond est franche et qui correspondent en général à des lésions bénignes.

b) Masqué : Un contour masqué est un contour qui est caché par le tissu normal adjacent. Ce terme est employé pour caractériser une masse circonscrite dont une partie du contour est cachée.

c) Microlobulé : c'est-à-dire qu'il comporte de petites ondulations.

d) Indistinct Dans ce cas, le contour est mal défini. Ce caractère indistinct (le contraire de circonscrit) peut correspondre à une infiltration.

e) Spiculé: c'est-à-dire comportant des structures filiformes qui rayonnent en s'éloignant du centre de l'opacité, et qui sont hautement suggestives de malignité.

2.3.2.3 La densité

L'aspect du sein normal est très variable d'une femme à l'autre. Le facteur le plus remarquable est la grande variabilité de la densité radiologique de l'aire mammaire. Wolfe est le premier qui a établi une relation entre la densité du tissu mammaire et le risque de développer un cancer [24]. La classification BIRADS de l'ACR (American College of Radiology) définit 4 classes de la composition du sein (figure 2.9).

a) Stade 1 : Le sein est presque entièrement graisseux et homogène, radio transparent et facile à lire (moins de 25 % de la glande mammaire).

b) Stade 2 : Il y a des opacités fibro glandulaires dispersées. Le sein est graisseux et hétérogène (approximativement 25 à 50 % de la glande mammaire).

c) Stade 3 : Le tissu mammaire est dense et hétérogène (approximativement 51 à 75 % de la glande mammaire).

d) Stade 4 : Le tissu mammaire est extrêmement dense et homogène. La mammographie est alors difficile à interpréter puisque la densité peut masquer une lésion (plus de 75 % de la glande mammaire) (figure 2.9).

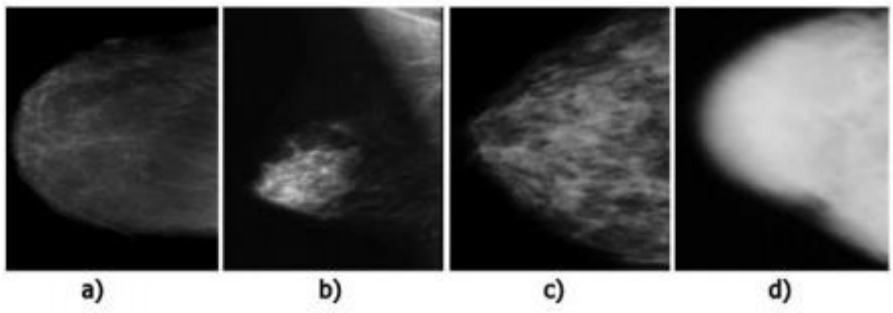

***Figure 2.9** : Densité mammaire selon le lexique BIRADS [20].*

2.4 La classification des anomalies mammaires

Il est important d'adopter un lexique standard et une classification commune afin de fournir aux radiologues une description claire et précise des lésions mammaires. L'étude morphologique de ces lésions a fait l'objet de plusieurs classifications à savoir la classification de Le Gal [26], de Lanyi [21] et de BIRADS [22]. Les classifications les plus connues et les plus pratiquées sont celles de Le Gal et de BIRADS.

2.4.1 La classification de BI-RADS

La classification BI-RADS est utilisée par les radiologistes lors de la mammographie, de l'échographie et de l'IRM pour définir les anomalies et permettre de savoir ce qui doit être fait par la suite, soit retour au dépistage, suivi rapproché ou biopsie. La dernière version classifie les lésions en sept catégories des images mammographiques en fonction du degré de suspicion de leur caractère pathologique (en dehors des images construites et des variantes du normal) :

BI-RADS 0

L'évaluation est incomplète et nécessitant des examens d'imagerie complémentaires. Cette catégorie est presque toujours utilisée en situation de dépistage mais rarement en situation diagnostique. Les recommandations sont : comparaison avec clichés antérieurs, clichés complémentaires, échographie etc. C'est une classification d'attente et le radiologue doit indiquer dans quelle mesure la poursuite des investigations.

BI-RADS 1

Mammographie normale. Les seins sont symétriques et il n'y a pas d'opacité, de distorsion architecturale ou de calcification suspecte. En principe cela ne pose aucun problème si les seins sont radio-transparents. Mais quelle est la certitude en cas de seins denses (type 3 ou 4) ? .Les solutions possibles résident dans la comparaison avec les mammographies anciennes, l'examen clinique ce qui renvoie à la question plus générale de l'intérêt de l'échographie systématiquement associée à l'exploration des seins denses.

BI-RADS 2

Constatations bénignes : « Cela correspond pratiquement à une mammographie négative mais le radiologue désire décrire une anomalie dont les caractères radiologiques sont caractéristiques et qui peuvent être étiquetées avec beaucoup de confiance dans l'interprétation. » Le risque d'association à un cancer est insignifiant en pratique et il n'y a pas lieu de poursuivre les investigations.il existe des anomalies bénignes ne nécessitant ni surveillance ni examen complémentaire :

- opacité ronde avec macro-calcifications (adénofibrome ou kyste) (figure2.10);

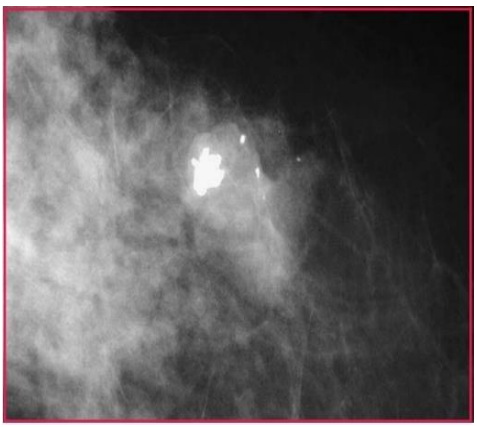

***Figure 2.10** : Aspect typique de fibroadénome partiellement calcifié : BI-RADS 2 [25].*

- ganglion intramammaire ;
- opacités rondes correspondant à un/des kystes typiques en échographie ;
- images de densité graisseuse ou mixte (lipome, hamartome, galactocèle, kyste huileux)
- cicatrices connues et calcifications sur matériel de suture ;
- macro-calcifications sans opacité (adénofibrome, kyste, adiponécrose, ectasie canalairesécrétante, calcifications vasculaires, etc.) ;
- micro-calcifications annulaires ou arciformes, semi-lunaires, sédimentées, rhomboédriques (calcifications d'aspect carré ou rectangulaire de face, losangiques ou trapézoïdales de profil, à étudier sur des agrandissements) (figure 2.11).

***Figure 2.11** : Micro calcifications bénignes : lait calcique : BI-RADS 2 [25].*

– calcifications cutanées et calcifications punctiformes régulières diffuses.

BI-RADS 3

Il existe une anomalie probablement bénigne pour laquelle une surveillance à court terme est conseillée :

- micro-calcifications rondes ou punctiformes régulières ou pulvérulentes, peu nombreuses, en petit amas rond isolé ;
- petit(s) amas rond(s) ou ovale(s) de calcifications amorphes, peu nombreuses, évoquant un début de calcification d'adénofibrome ;
- opacité(s) bien circonscrite(s), ronde(s), ovale(s) ou discrètement polycyclique(s) sans micro-lobulation, non calcifiée(s), non liquidiennes en échographie ;
- asymétrie focale de densité à limites concaves et/ou mélangée à de la graisse.

BI-RADS 4

Il existe une anomalie indéterminée ou suspecte qui indique une vérification histologique :

- micro-calcifications punctiformes régulières nombreuses et/ou groupées en amas aux contours ni ronds, ni ovales ;
- micro-calcifications pulvérulentes groupées et nombreuses ;
- micro-calcifications irrégulières, polymorphes ou granulaires, peu nombreuses
- image(s) spiculée(s) sans centre dense ;
- opacité(s) non liquidienne(s) ronde(s) ou ovale(s) aux contours lobulés, ou masqués, ou ayant augmenté de volume ;
- distorsion architecturale en dehors d'une cicatrice connue et stable ;
- asymétrie(s) ou surcroît(s) de densité localisé(s) à limites convexes ou évolutif(s).

BI-RADS 5

Il existe une anomalie évocatrice d'un cancer :

- micro-calcifications vermiculaires, arborescentes ou micro-calcifications irrégulières, polymorphes ou granulaires, nombreuses et groupées :
- groupement de micro-calcifications quelle que soit leur morphologie, dont la topographie est galactophorique ;
- micro-calcifications associées à une anomalie architecturale ou à une opacité ;
- micro-calcifications groupées ayant augmenté en nombre ou micro-calcifications dont la morphologie et la distribution sont devenues plus suspectes ;
- opacité mal circonscrite aux contours flous et irréguliers ;
- opacité spiculée à centre dense.

BI-RADS 6

Résultat de biopsie connu : malignité prouvée. Une action appropriée doit être entreprise. Utilisée dans le bilan d'extension et préthérapeutique de lésions malignes biopsées.

2.4.2 La classification de LeGal

En 1976, Le Gal du Service de Radio diagnostic de l'institut Curie à Paris, a conçu la classification dite de Le Gal [26]. Elle décrit cinq types morphologiques qui ont une valeur prédictive de malignité croissante (Tableau2.1):

Type 1	Mcs annulaires, arciformes ou polyédriques. Risque de cancer du sein quasi nul.
Type 2	Mcs rondes et de tailles variables. Risque de carcinome : 22%.
Type 3	Mcspoussiéreuses**,**pulvérulentes . Risque de cancer : 36%.
Type 4	Mcsirréguliéres associées à un risque de cancer : 56%.
Type 5	Mcs vermiculaires ou branchées. Risque de carcinome : 90%.

***Tableau 2.1** : Classification de Le Gal.*

Cette ancienne classification a l'avantage d'être simple. Toutefois, son défaut principal est qu'elle se base uniquement sur les microcalcifications et n'intègre pas d'autres paramètres tels que :

- L'étude morphologique des masses.
- La disposition des microcalcifications.
- L'étude des distorsions architecturales.
- Le comportement du radiologue vis-à-vis de chaque cas.

2.5 Conclusion :

Dans ce chapitre, nous avons étudié les outils de détection et diagnostic de cancer, les techniques qui nous permet de faire, l'échographie, leurs avantages et inconvénients. Prise en compte les techniques de classification ; BIRADS pour classifier les lésions mammaires en bénignes-malignes, qui nous permet de faire étudier les masses et les micros classification.

Chapitre 3

Les algorithmes génétiques

3.1 Introduction

Au cours de ce chapitre, nous avons présenté les algorithmes génétiques et ses paramètres, passons par ses principes, ensuite les quatre phases principales qui sont les opérateurs génétiques. Enfin nous présentons les avantages et les inconvénients.

3.2 Présentation des Algorithmes génétiques

Les algorithmes génétiques (AG) sont des méthodes d'optimisation stochastiques est actuellement bien connue. Les algorithmes évolutionnaires sont des méta-heuristiques qui s'inspirent des mécanismes d'évolution darwinienne des populations biologiques.

Les AGs ont été utilisés pour la première fois en 1950 par des biologistes dans le but de simuler l'évolution des organismes. Les algorithmes génétiques proposés par Holland [30] ont été utilisés avec un succès croissant en optimisation combinatoire. Un AG opère sur une population d'individus codés par des chaînes de symboles appelées chromosomes. Ces chaînes sont munies d'une fonction d'évaluation appelée fonction de fitness, qui correspond à une mesure d'adaptation au milieu.

Un AG procède par une multitude d'itérations où chaque itération consiste à tirer au sort deux parents selon une distribution favorisant les individus les plus adaptés. Un opérateur de croisement combine ensuite les deux chromosomes parents pour construire un ou deux enfants, pouvant à leur tour être modifiés aléatoirement par un opérateur de mutation. Les enfants servent à construire la génération suivante ou remplacent directement des individus de la population. Le processus est répété jusqu'à ce qu'un critère d'arrêt, défini par l'utilisateur, soit vérifié. Les sections suivantes présentent les caractéristiques essentielles des AG : les chromosomes, leur évaluation, les méthodes de croisements et les mutations. Ces caractéristiques peuvent se décliner en différents AG selon l'implémentation de la population initiale, la gestion des générations et les critères d'arrêt.

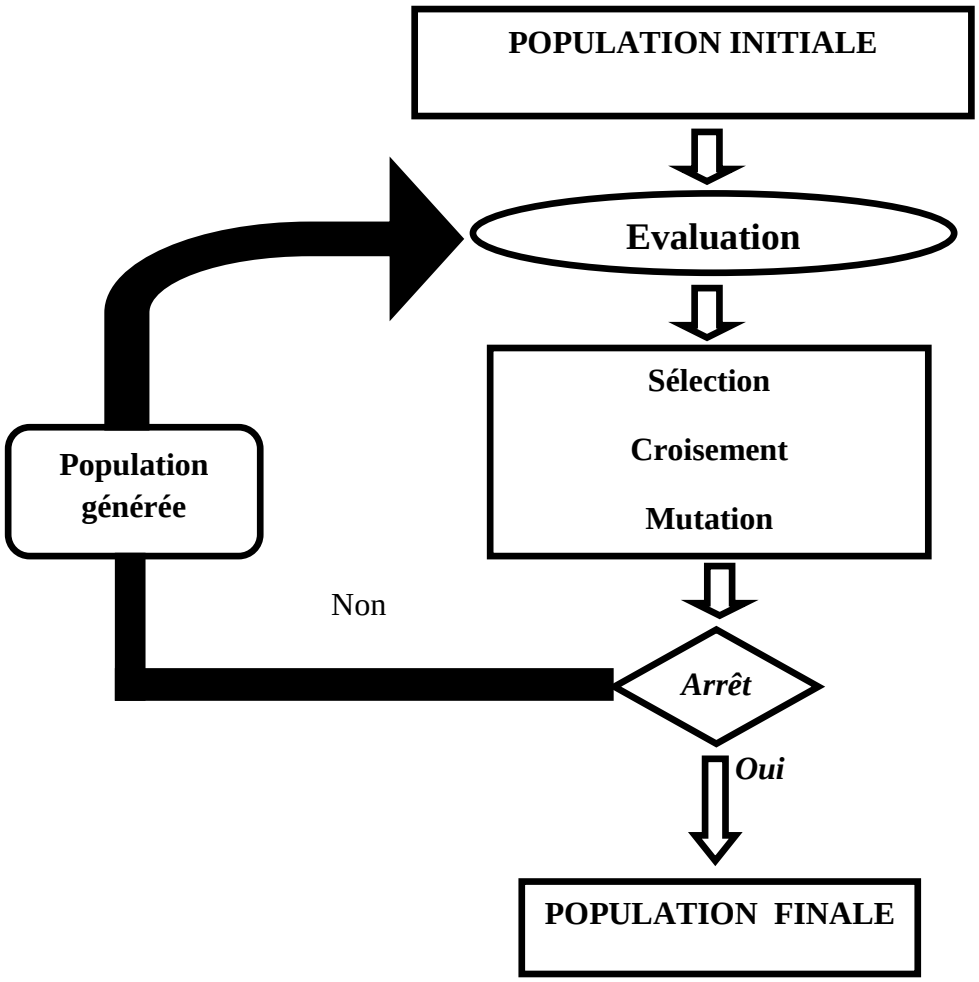

***Figure 3.1** : Schéma général d'un algorithme génétique.*

3.2.1 Principe

Les algorithmes génétiques sont inspirés des mécanismes de la sélection naturelle et de la génétique. Ils utilisent les principes de la survie des individus considérés comme les plus forts ou les mieux adaptés à l'environnement. Il s'agit alors de combiner les points forts de chaque individu pour en créer de nouveaux de manière à ce que leur efficacité soit meilleure. Avec les AG on cherche à optimiser une fonction (*objectif*) donnée dans un espace de recherche, celui des individus. Pour l'optimiser, on définit une fonction d'évaluation (*fitness*) reliée à cette fonction *objectif* et appliquée sur chaque individu ou chromosome. En général, le fonctionnement d'un AG est basé sur les phases suivantes (Figure 3.1) :

1. **Initialisation** : Générer aléatoirement une population initiale de taille *N* chromosomes.
2. **Evaluation**: Evaluer chaque individu de la population par la fonction d'évaluation appropriée au problème (fonction de fitness).

3. **Reproduction** : Créer une nouvelle population de *N* chromosomes par l'utilisation d'une méthode de sélection appropriée et l'application d'opérateurs génétiques (croisement et mutation) sur certains chromosomes au sein de la population courante.
4. **Retour :** à la phase 2(Evaluation) tant que la condition d'arrêt du problème n'est pas satisfaite.

3.2.2 Type de chromosome et évaluation

Les chromosomes traduisent directement les différentes solutions du système. Pour une solution réalisable, un chromosome est la concaténation de gènes, c'est-à-dire, un ensemble composé des paramètres de forme et des angles de rotation, des translations et des coefficients de mise à l'échelle. Cet ensemble est noté { *m,n1,n2,n3,M,N1,N2,N3,tx,ty,tz,* φ,θ,ψ,*sx,sy,sz*}.

3.3 Caractéristiques d'algorithme génétique

Les algorithmes génétiques se caractérisent par certains aspects à savoir : le codage des paramètres du problème, la fonction d'évaluation servant à sélectionner les chromosomes parents, et le hasard qui joue un rôle important dans l'évolution de chromosomes de génération en génération, dans ce que suite les aspects:

3.3.1 Codage des données

Le AG initial de Holland [1] utilisait des chromosomes binaires. Ce type de codage a pour avantage la création d'opérateurs de croisement et de mutation simples par des opérations de masque et des opérations booléennes. Cependant, ce type de codage n'est pas toujours adapté à certains types de problèmes comme le montre l'exemple suivant : deux éléments voisins en terme de distance de Hamming (qui représente le nombre de bits dont diffèrent deux nombres binaires) ne sont pas forcément codés comme deux éléments proches dans l'espace de recherche. Ainsi, les nombres binaires 10000000 et 00000000 ont, certes, une distance de Hamming de 1, qui est une distance faible, mais représentent des valeurs réelles très éloignées (128 et 0). C'est pour cette raison que nos chromosomes sont codés par des nombres réels.

3.3.2 Fonction d'évaluation ou La fitness

La fitness, ou encore fonction d'évaluation, quantifie la qualité de chaque chromosome par rapport au problème. Elle est généralement utilisée pour sélectionner les chromosomes pour la reproduction. Les chromosomes ayant une bonne qualité ont alors plus de chances d'être

sélectionnés pour la reproduction, faisant en sorte que la prochaine génération de la population hérite de leur matériel génétique. La fonction d'évaluation produit la pression qui permet de faire évoluer la population de L'AG vers des individus de meilleure qualité.

Dans un problème d'identification, l'objectif est la minimisation d'une fonction coût, ou les algorithmes d'évolution maximisent une fonction d'adaptation, qui doit être une mesure de profit positive. Aussi est-il nécessaire de transformer la fonction. C'est ici très facile puisque la fonction coût est toujours positive. Nous choisirons donc une fonction d'adaptation de la forme :

$$f_{a(x)}=\frac{1}{f(x)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

3.3.3 Rôle hasard

Les algorithmes génétiques utilisent des règles de transition probabilistes plutôt que déterministes pour guider leur recherche. Le choix des chromosomes à perturber est réalisé de façon probabiliste. Dans le processus de croisement, le lieu de croisement est choisi aléatoirement à l'intérieur du chromosome. De même, le gène devant subir une mutation à l'intérieur du chromosome est choisi selon une certaine probabilité .le hasard occupe donc une place importante dans le fonctionnement des algorithmes génétiques.

3.4 Les opérateurs génétiques

Les opérateurs génétiques travaillent directement sur les individus composant la population. Nous en dénombrons trois principaux : opérateurs d'initialisation**,** opérateur de sélection, de croisement et de mutation. Si le principe de chacun de ces opérateurs est facilement compréhensible, il est toutefois difficile d'expliquer l'importance isolée de chacun de ces opérateurs dans la réussite de l'AG. Ils permettent, notamment dans le cas mutation ou du croisement, de créer des nouvelles solutions.

3.4.1 Les Opérateurs d'initialisation

Cet opérateur est utilisé pour générer la population initiale de l'algorithme génétique .la population initiale doit contenir des chromosomes qui soient bien répartis dans l'espace des solutions pour fournir à l'algorithme génétique un matériel génétique varié .la façon la plus simple est de générer aléatoirement les chromosomes.

3.4.2 Opérateur de la sélection

Cet opérateur est chargé de définir quels seront les individus de *P* qui vont être dupliqués dans la nouvelle population *P'* et vont servir de parents (application de l'opérateur de croisement). Soit *n* le nombre d'individus de *P*, on doit en sélectionner *n/2* (l'opérateur de croisement nous permet de repasser à *n* individus). Cet opérateur est peut-être le plus important puisqu'il permet aux individus d'une population de survivre, de se reproduire ou de mourir. En règle générale, la probabilité de survie d'un individu sera directement reliée à son efficacité relative au sein de la population.

On trouve essentiellement quatre types de méthodes de sélection différentes :

• La méthode de la "loterie biaisée" (roulette wheel),

• La méthode "élitiste",

• La sélection par tournois,

• La sélection par rang.

a) La loterie biaisée ou roulette wheel

Pour chaque individu, on reproduit sur un disque de périmètre un arc de cercle proportionnel à sa *fitness relative*, puis on fait tourner la roulette aléatoirement et on reproduit l'individu sélectionné ; on recommence *n* fois pour obtenir la nouvelle génération. Le nombre de descendant est ainsi statistiquement proportionnel à sa *fitness*. Cependant, plus la taille de la population est faible, plus la sélection sera biaisée à cause du petit nombre de tirages effectués.

fitness relative d'un individu se calcule en rapportant sa fitness à la somme des fitness de tous les individus :

$$f_{relative}(ind^i)=\frac{f(ind^i)}{\sum_{j\in[1,n]} f(ind^i)} \qquad \ldots\ldots\ldots\ldots (2)$$

Cette méthode, bien que largement répandue, à pas mal d'inconvénients :

• En effet, elle a une forte variance. Il n'est pas impossible que sur n sélections successives destinées à désigner les parents de la nouvelle génération *P*', la quasi-totalité, voire pire la totalité des n individus sélectionnés soient des individus ayant une fitness vraiment mauvaise et donc que pratiquement aucun individu voire aucun individu a forte fitness ne fasse partie des parents de la nouvelle génération. Ce phénomène est bien sûr très dommageable car cela

va complètement à l'encontre du principe des algorithmes génétiques qui veut que les meilleurs individus soient sélectionnés de manière à converger vers une solution la plus optimale possible.

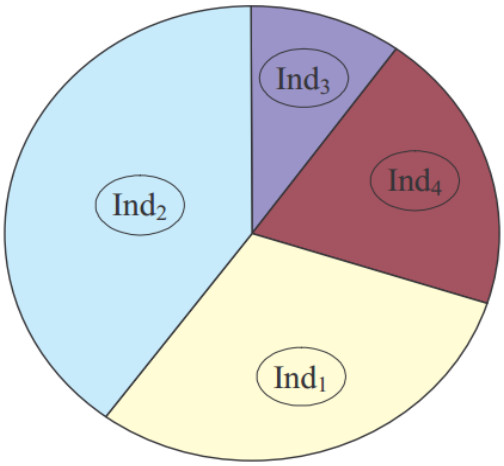

***Figure 3.2** : Méthode de la sélection loterie baisée.*

• A l'inverse, on peut arriver à une domination écrasante d'un individu "localement supérieur". Ceci entrainant une grave perte de diversité. Imaginons par exemple qu'on ait un individu ayant une fitness très élevée par rapport au reste de la population, disons dix fois supérieure, il n'est pas impossible qu'après quelques générations successives on se retrouve avec une population ne contenant que des copies de cet individu. Le problème est que cet individu avait une fitness très élevée, mais que cette fitness était toute relative, elle était très élevée mais seulement en comparaison des autres individus. On se retrouve donc face à problème connu sous le nom de "convergence prématurée; l'évolution se met donc à stagner et on atteindra alors jamais l'optimum, on restera bloqué sur un optimum local. Il existe certaines techniques pour essayer de limiter ce phénomène, comme par exemple le "scaling", qui consiste à effectuer un changement d'échelle de manière à augmenter ou diminuer de manière forcée a fitness d'un individu par rapport à un autre selon leur écart de fitness. Malgré tout, il est conseillé d'opter plutôt pour une autre méthode de sélection.

b) La méthode élitiste

Cette méthode consiste à sélectionner les *n* individus dont on a besoin pour la nouvelle génération *P'* en prenant les n meilleurs individus de la population *P* après l'avoir triée de manière décroissante selon la fitness de ses individus. Il est inutile de préciser que cette méthode est encore pire que celle de la loterie biaisée dans le sens où elle amènera à une convergence prématurée encore plus rapidement et surtout de manière encore plus sûre que la

méthode de sélection de la loterie biaisée ; en effet, la pression de la sélection est trop forte, la variance nulle et la diversité inexistante, du moins le peu de diversité qu'il pourrait y avoir ne résultera pas de la sélection mais plutôt du croisement et des mutations. Là aussi il faut opter pour une autre méthode de sélection.

c) La sélection par tournoi

Cette méthode est celle avec laquelle on obtient les résultats les plus satisfaisants. Le principe de cette méthode est le suivant : on effectue un tirage avec remise de deux individus de *P*, et on le fait "combattre". Celui qui a la fitness la plus élevée l'emporte avec une probabilité *p* comprise entre 0.5 et 1. On répète ce processus n fois de manière à obtenir les n individus de *P'* qui serviront de parents. La variance de cette méthode est élevée et le fait d'augmenter ou de diminuer la valeur de *p* permet respectivement de diminuer ou d'augmenter la pression de la sélection.

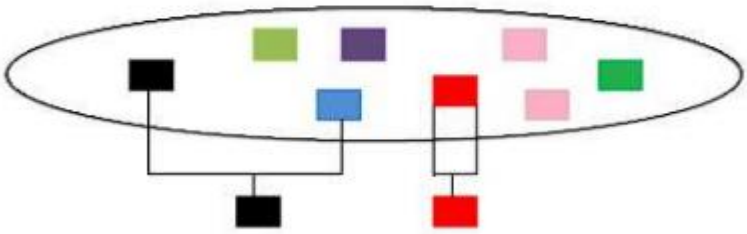

Figure 3.3 *: Le Tournoi.*

d) Sélection par rang

Elle consiste à attribuer à chaque individu son classement selon l'ordre donné par la fonction d'évaluation qui quantifie l'adaptation de l'individu comme solution du problème. Le plus mauvais individu prendra 1 comme rang, par contre le meilleur aura le rang N, où N est le nombre d'individus de la population. La probabilité de sélection d'un individu *xi* devient :

$$P_{sel\ (x_i)} = \frac{Rang(x^i)}{\sum_{j=1}^{N} Rang(x^i)} \qquad \ldots\ldots\ldots\ldots\ldots (3)$$

Avec cette méthode de sélection, tous les chromosomes ont une chance d'être sélectionnés. Cependant, cette méthode conduit à une convergence plus lente des individus de la population vers la bonne solution. Cela est dû au fait que les meilleurs chromosomes ne différent pas toujours énormément des plus mauvais.

3.4.3 Opérateur de croisement

L'opérateur de croisement combine le matériel de un ou plusieurs parents pour obtenir un ou plusieurs enfants.il existe différents types de croisement, nous allons brièvement présenter les trois principaux dont une illustration est donnée dans la figure3.4.Le croisement un point détermine aléatoirement un point de coupure et échange la deuxième partie des deux parents. Le croisement deux points (qui peut être étendu à x points) possède 2 points (ou x) de coupures qui sont déterminés aléatoirement. Enfin le croisement uniforme échange chaque bit avec une probabilité fixé à 1/2.

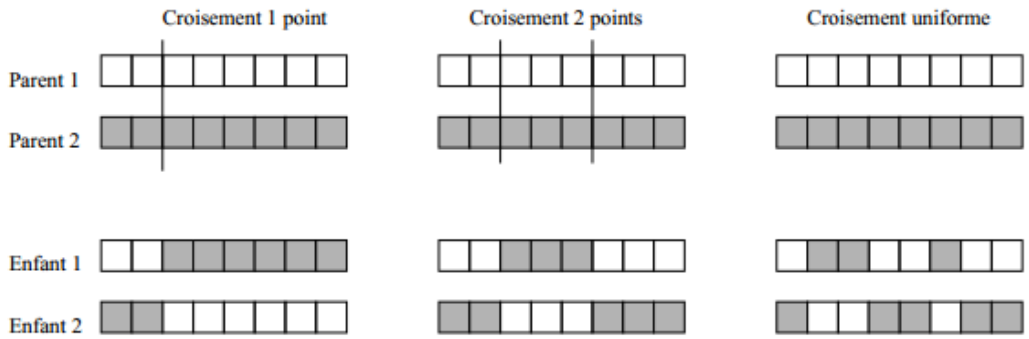

***Figure 3.4** : les exemples du croisement.*

3.4.4 Opérateur de mutation

Cet opérateur consiste à changer la valeur allélique d'un gène avec une probabilité *pm* très faible, généralement comprise entre 0.01 et 0.001. On peut aussi prendre *pm = 1 / lg* où l*g* est la longueur de la chaîne de bits codant notre chromosome. Une mutation consiste simplement en l'inversion d'un bit (ou de plusieurs bits, mais vu la probabilité de mutation c'est extrêmement rare) se trouvant en un locus bien particulier et lui aussi déterminé de manière aléatoire; on peut donc résumer la mutation de la façon suivante :On utilise une fonction censée nous retourner *true* avec une probabilité *pm* .

Pour chaque locus **faire**

Faire appel à la fonction
Si cette fonction nous renvoie *true***alors**

on inverse le bit se trouvant à ce locus
FinSi
FinPour

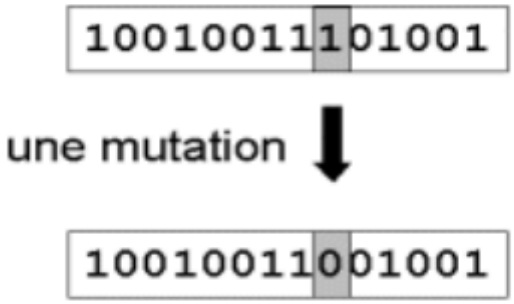

***Figure 3.5** : Une mutation.*

L'opérateur de mutation modifie donc de manière complètement aléatoire les caractéristiques d'une solution, ce qui permet d'introduire et de maintenir la diversité au sein de notre population de solutions. Cet opérateur joue le rôle d'un "élément perturbateur", il introduit du "bruit" au sein de la population.

La mutation est traditionnellement considérée comme un opérateur marginal bien qu'elle confère en quelque sorte aux algorithmes génétiques la propriété d'ergodicité (i.e. tous les points de l'espace de recherche peuvent être atteints). Cet opérateur est donc d'une grande importance. Il a de fait un double rôle : celui d'effectuer une recherche locale et/ou de sortir d'une trappe (recherche éloignée).

Parmi les stratégies de mutation utilisées en pratique [32] :

- **Mutation unipoint :**

Cette mutation se fait par altération d'une seule valeur sur le chromosome.

- **Mutation bipoints et multipoints :**

Cette mutation se fait par altération de plusieurs valeurs sur le chromosome.

- **Mutation par valeur :**

Avec ce type, la mutation se fait par transformation d'une valeur donnée en une autre valeur déterminée, sur tous les gènes de chromosome.

En effet, une mutation pouvant intervenir de manière aléatoire au niveau de n'importe quel locus, on a la certitude mathématique que n'importe quel permutation de notre chaîne de bits peut apparaître au sein de la population et donc que tout point de l'espace de recherche peut être atteint .On notera que la mutation règle donc le problème exposé après le croisement.

3.5 Les paramètres de dimensionnement

Le processus de algorithme génétique est guidé par un certain nombre de paramètres fixés à l'avance .Ces paramètres sont les suivants :

- La taille de la population *N*, et la longueur du codage de chaque chromosome *l. Si N* est trop grand, *le* temps de recherche par l'algorithme devient important .Si N trop petit, la population peut converger trop rapidement vers un mauvais individu.
- La probabilité de croisement : elle dépend de la forme de la fonction de fitness. Plus elle élevée, plus la population subit des changements importants. Les valeurs généralement admis sont comprises entre 0.5 et 0.9.
- La probabilité de mutation : ce taux est généralement faible puisqu'un taux élève risque de conduire à une solution sous-optimale, et à la perte de la population originale.
- Le nombre de générations peut également être défini a priori comme critère d'arrêt.

3.6 Critère d'arrêt

C'est une des principales difficultés de ce type d'algorithme. A moins de fonctionner en mode interactif, il est bon de disposer d'un mécanisme automatique coupant le processus. Le moyen le plus simple est de fixer à l'avance le nombre de générations. Dans la mesure où seule une estimation de la zone contenant la solution est recherchée, ce test sera suffisant pour notre étude. Le suivi de l'adaptation du meilleur membre de la population peut également fournir un critère d'arrêt : on décide de s'arrêter si aucune amélioration significative n'est observée. L'examen du contenu génétique de la population courante permet également d'envisager un critère. En effet, si tous les individus se ressemblent, l'optimisation ne progresse qu'à l'aide de l'opérateur de mutation, ce qui est très peu efficace.

On distingue deux grandes catégories de critères d'arrêt :

– un critère *statique* est généralement basé sur les ressources matérielles disponibles (temps CPU, nombre d'itérations ou d'évaluations de la fonction objectif) et connues a priori ;

– un critère *dynamique* fait référence à la qualité de la solution (suffisamment proche d'un optimum connu a priori) ou à la fin de convergence (nombre d'itérations consécutives sans améliorer la meilleure solution connue).

3.7 Les avantages et les inconvénients des AG

3.7.1 Les avantages

L'algorithme génétique présente plusieurs points forts comme :

- ✓ Le fait d'utiliser seulement l'évaluation de la fonction objectif sans se soucier de sa nature. En effet nous n'avons besoin d'aucune propriété particulière sur la fonction à optimiser (continuité, convexité, dérivabilité, etc.), ce qui lui donne plus de souples et un large domaine d'applications.
- ✓ Génération d'une forme de parallélisme en travaillant sur plusieurs points en même temps (population de taille N).
- ✓ L'utilisation des règles de transition probabilistes (probabilités de croisement et de mutation).
- ✓ Pour les mêmes raisons un algorithme génétique est dans l'idéal totalement indépendant de la nature de problème et de la fonctionnelle à optimiser, car il ne sert que des valeurs d'adaptation, qui peuvent être très différentes des valeurs de la fonction à optimiser, mêmes si elles sont calculées à partir de cette dernière.
- ✓ Potentiellement les algorithmes génétiques explorent tous l'espace des points en même temps, ce qui limite les risques de tomber dans des optimums locaux. Les algorithmes génétiques ne se servent que des valeurs de la fonctionnelle pour optimiser cette dernière, il n'y a pas besoin d'effectuer de coûteux et parfois très complexes calculs.
- ✓ Les algorithmes génétiques présentent une grande robustesse c'est-à-dire une grande capacité à trouver les optimums globaux des problèmes d'optimisation.

3.7.2 Les inconvénients

- ❖ Les algorithmes génétiques ne sont encore actuellement pas très efficaces en coût(ou vitesse de convergence), vis-à-vis de méthodes d'optimisation plus classique ;
- ❖ Parfois les algorithmes génétiques convergents très vite vers un individu particulier de la population dans la valeur d'adaptation est très élevée ;
- ❖ le respect de la contraint de domaine par la solution codées sous forme de chaîne de bits pose parfois problème, il faut bien choisir le codage, voire modifier les opérateurs ;
- ❖ L'utilisation d'un algorithme génétique ne garantit pas le succès de l'optimisation ;
- ❖ En pratique l'efficacité d'un AG dépend souvent de la nature du problème d'optimisation .selon les cas de choix des opérateurs et des paramètres seront souvent critiques, mais aucune théorie générale ne permet de connaître avec certitude la bonne paramétrisation, il faudra faire plusieurs expériences pour s'en approcher [31].

3.8 Conclusion

Dans le cadre de la sélection des caractéristiques(les caractéristiques sont calculées précédemment) des masses mammographie, nous avons crié un algorithme génétique passons par les étapes principales pour atteinte à une objective correspond le meilleur individu.

Chapitre 4

Conception et réalisation

4.1 Introduction

La classification du cancer du sein reste est une tâche difficile à résoudre à cause de variabilité des clichés mammographies, plusieurs formes de cancer du sein, plusieurs types du cancer.

Dans ce chapitre, nous proposons une approche génétique pour la sélection de caractéristiques des masses mammographique, cette approche s'inspire globalement de l'approche du médecin lors de l'examen radiologique comme c'était convenu dans le système d'aide à la rédaction des comptes rendus BI-RADS (Breast Imaging Reporting And Data System), ce dernier permet de décrire les anomalies rencontrées en mammographie.

4.2 Environnement et outils de développement

4.2.1 Plateformes utilisées

Notre application a été implémentée sous Windows 10. Les clichés mammographique sont décompressés sous linux elementary OS.

4.2.2 Langage et environnement

Langage de programmation JAVA avec environnement " Eclipse" qui est principalement écrit en Java.

Java est un langage de programmation et une plate-forme informatique qui a été crée par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable.

La version d'Eclipse utilisée est "Luna" sortie de 07 juillet 2014 est composée de 79 projets cette dernière apporte quelques nouveautés majeur : et le déploiement d'une version d'Eclipse avec les configurations associées au sein d'une équipe développement.

- L'explorateur de projets peut désormais afficher des projets inclus dans d'autres projets de manière hiérarchique.

- Une partie des outils de développement a été intégrée directement sur les "update-sites".

4.2.3 La base d'images utilisées

Pour mesurer les performances de notre système de classification on a besoin d'un référentiel dans le domaine d'aide au diagnostic du cancer du sein, il existe plusieurs bases d'images : MIAS (Mammographic Image Analysis So ciety), DDSM (Digital Database for Screening Mammography), AMDI (Indexed Atlas of Digital Mammograms) et WDBC (Wisconsin DiagnosisBreast Cancer).

Nous avons utilisé un sous-ensemble de la base DDSM. Les types de tumeurs pris en compte dans notre système sont des masses bénigne et maligne. DDSM : La base de données Marathon de l'université de la Floride du Sud. Une description de cette base a été effectuée par "American college of Radiology" dans le lexique de BI-RADS (Breast Imaging Reporting and Data System).

La base contient 2604 dossiers de patients. Chaque dossier de patiente est composé de : -1 fichier .ics décrivant en format ASCII, les informations générales d'un dossier de patient.

-4 fichiers images .LJPEG (LOSSLESS JPEG) des radios numérisées.

-Chaque radio présente un angle de vue du sein : Left-CC, Left-MLO, Right-CC, Right-MLO (CC : Cranio Caudal ,MLO : Medio Latral Oblique).

-Pour chaque radio présentant une ou des zones anormales, est associé un fichier .OVERLAY en format ASCII, décrivant une anomalie du sein.

-fichier image .16-PGM regroupant les 4 radios et présentant un aperçu rapide pour la visualisation d'un dossier de patient.

L'avantage majeur de la base DDSM est qu'elle emploie le même lexique standardisé, Par l'American College of Radiology dans le BIRADS.

Dans cette phase des informations liées aux clichés mammographiques et aux anomalies sont extraites automatiquement à partir des fichiers .ics et .overlay, ces derniers nous offres des informations sur les images et sur les anomalies.

4.2.3.1 Description du fichier .ics

Le fichier .ics contient des informations importantes telles que la date de l'étude, l'âge du patient, la date de la numérisation des films, le type de numériseur utilisé, une densité du tissu mammaire, la taille de chaque fichier image, le nombre de bits par pixel, la résolution de la numérisation, etc .La figure suivante illustre un exemple de ce fichier.

```
ics_version 1.0
filename A-1141-1
DATE_OF_STUDY 19 4 1991
PATIENT_AGE 53
FILM
FILM_TYPE REGULAR
DENSITY 4
DATE_DIGITIZED 17 6 1998
DIGITIZER HOWTEK 43.5
SEQUENCE
LEFT_CC LINES 5641 PIXELS_PER_LINE 2671 BITS_PER_PIXEL 12
RESOLUTION 43.5 NON_OVERLAY
LEFT_MLO LINES 6001 PIXELS_PER_LINE 2821 BITS_PER_PIXEL 12
RESOLUTION 43.5 NON_OVERLAY
RIGHT_CC LINES 5896 PIXELS_PER_LINE 2761 BITS_PER_PIXEL 12
RESOLUTION 43.5 OVERLAY
RIGHT_MLO LINES 5971 PIXELS_PER_LINE 3256 BITS_PER_PIXEL 12
RESOLUTION 43.5 OVERLAY
```

Figure 4.1 *: Le fichier A-1141-1.ics.*

4.2.3.2 Description du fichier overlay

Les cas anormaux ont entre un et quatre fichiers overlay, ceux-ci dépendent du nombre d'images que le radiologue marque comme sans anormalités. Pour chaque anormalité, on a des informations sur le nombre d'anormalités, le type, la forme et les bords de la tumeur présente, le degré de suspicion, le degré de subtilité, le type de pathologie et enfin la description se termine par le code du contour de l'anormalité. Pour toutes les mammographies contenant une tumeur, le spécialiste a tracé un contour autour de la région tumorale, ce contour est analysé sous forme de chaîne à l'aide du code de Freeman. Les deux premiers chiffres de la chaîne représentent les coordonnées d'un pixel du contour respectivement la colonne et la ligne, sur lesquels on se base pour déterminer les coordonnées des autres pixels du contour.

```
TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC DISTRIBUTION
CLUSTERED
ASSESSMENT 4
SUBTLETY 2
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
BOUNDARY
850 2007  0 0 0 0 0 2 2 2 2 0 0 0 0 1 1 1 1 1 0 0 0 0 0 2 2 2
... 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 #
```

Figure 4.2: *Fichier A_1141_1.RIGHT_CC. Overlay.*

4.2.3.3 Conversion du LJPEG au LJPEG1

LJPEG nommé aussi JPEG-LS (souvent surnommé Lossless JPEG) est une norme de compression sans perte (donc réversible), basée sur l'algorithme LOCO (LOw COmplexity LOssless COmpression for Images) et évaluée par le Joint Photographic Experts Group, dont la notoriété est reconnue pour les formats de compression JPEG ISO/CEI 10918-1 et JPEG 2000.

Dans JPEG-LS la compression est réalisée par la combinaison d'un codage adaptatif (extension des codes de Golomb) avec un codeur entropique proche du codeur de Huffman pour les zones à faible entropie. L'image est stockée étendue en tant que données binaires brutes dans un fichier qui ne contient pas d'informations d'entête. Les dimensions de l'image (hauteur et largeur) sont précisées dans le fichier "ics" pour ce cas et cette décompression est faite sous Linux.

4.3 La sélection génétique de caractéristiques

Dans ce travail nous avons proposé une approche génétique pour la sélection de caractéristiques, qui nous permettra par la suite de catégoriser les anomalies mammaires en deux classes (Malignes et bénignes), cette méthode dernier se caractérise par certains aspects à savoir : le codage des chromosomes, la fonction d'évaluation servant à sélectionner les chromosomes parents (fitness), et le hasard assure l'évolution de chromosomes de génération en génération.

Codage des chromosomes :

Notre approche est guidée par un certain nombre de paramètres fixés manuellement:

- La taille de la population *nbrChromosomes*, et Le nombre de générations est défini comme critère d'arrêt. Si ces derniers paramètres sont trop grands, le temps d'exploration génétique devient important .Sinon, la population peut converger trop rapidement vers un mauvais individu.
- La probabilité de croisement : elle dépend de la forme de la fonction de fitness. Plus elle élevée, plus la population subit des changements importants. La valeur est défini par 0,5.
- La probabilité de mutation : ce taux est généralement faible puisqu'un taux élève risque de conduire à une solution sous-optimale, et à la perte de la population originale.
- le nombre de caractéristiques de chaque chromosome *TailledeChromosome*.

4.3.1 Génération de la population initiale

C'est la première phase de la sélection consiste à générer aléatoirement une population ou génération initiale avec des individus défini pour chaque individu un numéro au hasard sous l'intervalle [0,59] qui sont l'ensemble des caractéristiques ou descripteurs.

4.3.2 L'évaluation

Pour évaluer les individus de la population initiale, chaque sous-ensemble de caractéristiques (chromosome) est évalué à travers la fonction de fitness. Le résultat de cette fonction est le taux de reconnaissance fourni par un perceptron multicouches (PMC), l'objectif principal du PMC est d'effectuer les phases d'apprentissage et d'évaluation, ce qui permet d'évaluer chaque chromosome.

4.3.3 Croisement

L'operateur de croisement combine deux ou plusieurs parents pour obtenir un ou plusieurs enfants, pour cela nous adoptons un croisement ou bien une fusion entre deux parents tel que chaque 2 parents produise 2 enfants et cette fusion présente une coupure ou point de croisement.

```
Pour i allant de 0 a taille.chromosome faire
Si  i % 2==0
                  Enfant1.ajouterparent1
                  Enfant2.ajouterparent2
Sinon
                  Enfant1.ajouterparent2
                  Enfant2.ajouterparent1
```

```
Population.ajouterEnfant1
Population.ajouterEnfant2
FinSi
FinSinon
```

Exemple croisement de deux chromosomes :

Supposons que les caractéristiques C1.1,C1.2,C1.3 et C1.4 sont les identifiants des qui composent le chromosome parent 1 et les caractéristiques C2.1, C2.2, C2.3 et C2.4 sont les identifiants des qui composent le chromosome parent 2.

Le principe de croisement des deux parents est illustré dans la figure 4.1

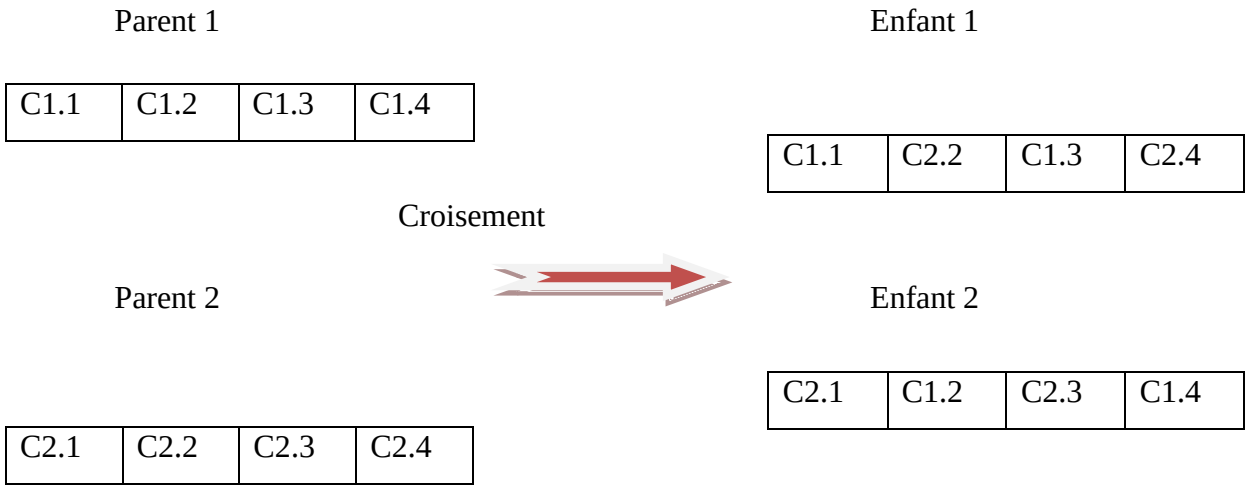

***Figure 4.3** : Croisement.*

4.3.4 Mutation

Il y a plusieurs stratégies utilisées en pratique : mutation unipoint, bipoints, multipoints, et mutation par valeur déterminé, pour notre algorithme nous avons effectué une mutation sélective appliquée seulement sur les caractéristiques répétées dans le même chromosome après la phase de croisement.

Supposant qu'après le croisement le caractéristique numéro 15 se répète deux fois dans le même chromosome comme le montre la figure 4.2.

Dans ce cas un nombre aléatoire est généré pour remplacer la caractéristique en état de répétition.

Enfant 1

20	30	57	15	15	46	16	14	22	13

20	30	57	33	15	46	16	14	22	13

***Figure 4.4**: Mutation.*

4.3.5 Critère d'arrêt

On distingue deux grandes catégories de critères d'arrêt :

– un critère *statique* est généralement basé sur les ressources matérielles disponibles (temps CPU, nombre d'itérations ou d'évaluations de la fonction objectif) et connues a priori ;

Et pour notre cas le taux de reconnaissance est défini comme un critère d'arrêt.

– un critère *dynamique* fait référence à la qualité de la solution (suffisamment proche d'un optimum connu a priori) ou à la fin de convergence (nombre d'itérations consécutives sans améliorer la meilleure solution connue).

Notre algorithme est défini par un nombre de générations limité.

4.4 Expérimentations

Dans ce projet nous avons effectué plusieurs expérimentations pour chercher le nombre de caractéristiques idéal qui caractérise la masse mammographique, et pour choisir les meilleures caractéristiques parmi les 59 caractéristiques utilisées par Sahtal et Benkirat [35].

4.4.1 L'évolution des taux de reconnaissance par rapport aux générations

Dans cette expérimentation nous avons manipulé quelques paramètres de l'algorithme génétique tel que :

- Le nombre de caractéristiques qui composent le chromosome a été fixé à 20.
- La taille de la population fixée : 20.

La figure 4.3 représente l'évolution des taux de reconnaissance des anomalies mammaires de la 1er jusqu'au la 100ème génération.

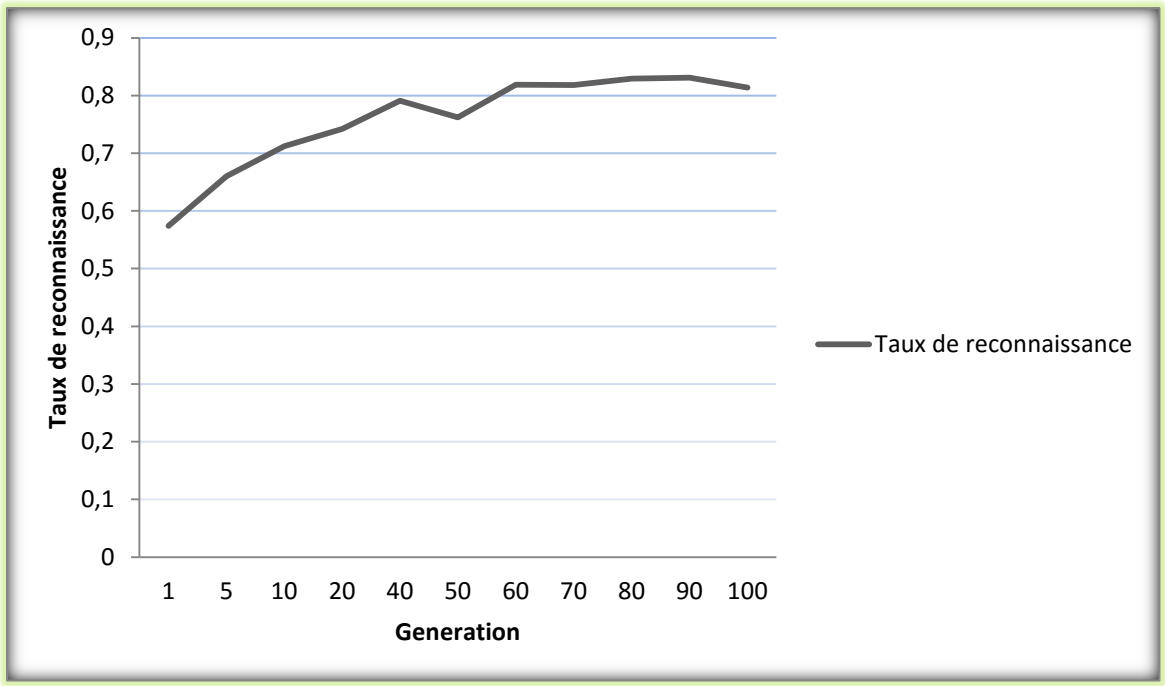

Figure 4.5 *: L'évolution des taux de reconnaissance.*

génération	**Taux de reconnaissance**
1	0,574
5	0.66
10	0.712
20	0.742
40	0.791
50	0.762
60	0.819
70	0.819
80	0.829
90	0.831
100	0.814

Tableau 4.1 : Meilleur taux de reconnaissance pour quelques générations.

Nous avons observé que la meilleure génération est la génération 90 avec un taux de reconnaissance 0.831, voir (Tableau 4.1).

4.4.2 L'évolution de taux de reconnaissance par rapport aux nombres de caractéristiques

Pour cette expérimentation nous avons fixé le nombre de générations à 30, et la taille de la population initiale à 20 chromosomes.

- Le nombre de caractéristiques pour chaque chromosome est varié de 1 à 52.

La figure suivante montre les taux de reconnaissances obtenus après 30 générations pour chaque taille de chromosome.

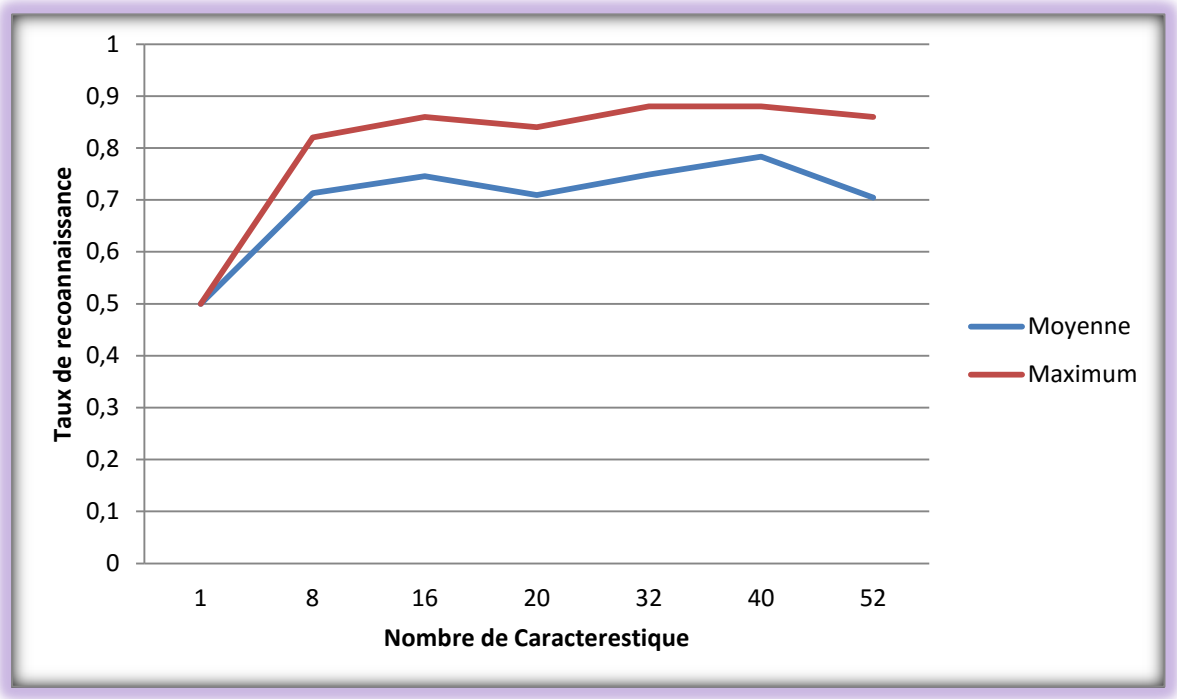

***Figure 4.6** : l'évolution de taux de reconnaissance par rapport au nombre de caractéristiques.*

La courbe en rouge représente les taux de reconnaissances maximaux de chaque expérimentation, et la courbe en bleu représente les valeurs moyennes de taux de reconnaissance de la population finale.

La meilleure configuration pour un meilleur taux de reconnaissance est : 40 caractéristiques pour un taux de reconnaissance moyenne de 0,783 et un taux maximum de 0.88 voir (Tableau 4.2).

Nombre de Caractéristique	Moyenne	Maximum
1	0,5	0,5
8	0,71297	0,82
16	0,7458	0,86
20	0,7094	0,84
32	0,7488	0,88
40	0,783	0,88
52	0,7046	0,86

***Tableau 4.2** : le taux de reconnaissance de chaque manipulation de caractéristiques.*

Les 40 descripteurs élus après 30 générations sont présenté dans le tableau suivant :

№	La caractéristique	№	La caractéristique
1	Corrélation	21	La compacité
2	Homogénéité	22	Mean variations
3	Variance	23	Variance variations
4	Contraste	24	Skewness variations
5	Énergie	25	Kurtosis variations
6	Probabilité maximale	26	Entropie variations
7	Somme moyenne	27	La courbure
8	Proéminence Cluster	28	La longueur radiale normalisée
9	Moyen absolu Écart	29	La moyenne de LRN (davg)
10	Minimum	30	La déviation standard de LRN
11	Maximum	31	L'Entropie (E)
12	Variance	32	La rugosité
13	Skewness	33	Le rapport de surface
14	Large différence emphasis (LDE)	34	Différence des déviations standard
15	Sharpness	35	La différence de l'entropie
16	Second Moment de DGD (SMG)	36	Le rapport de surface modifiée
17	Long distance emphasis for large différence (LDEL)	37	Moyenne de kurtosis
18	Gray Level Co-occurrence Matrix	38	Moyenne d'entropie
19	La circularité	39	Variation totale d'index de probabilité maximale
20	La rectangularité	40	Déviation standard d'entropie

***Tableau 4.3** : Les 40 caractéristiques les plus informatifs d'après notre expérimentation.*

4.5 Réalisation et Interface

L'interface de notre application (figure 4.5) est comme suit :

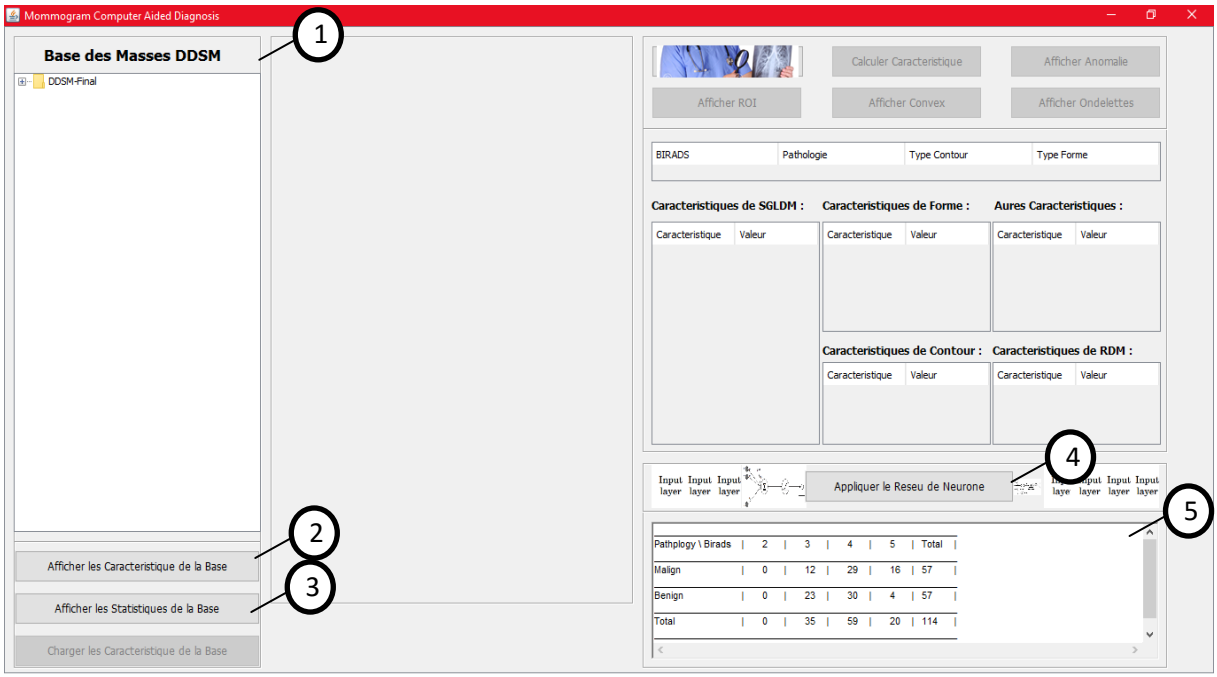

***Figure 4.7** : Interface principale de notre application.*

1. **Zone qui contient la base d'images DDSM**
2. **Afficher les caractéristiques de la base :** il sert à afficher les caractéristiques de tous les clichés mammaires de la base.
3. **Afficher les statistiques de la base :** afficher les statistiques du niveau de BIRADS ainsi que la pathologie bénigne/maligne de notre base DDSM.
4. **Appliquer le réseau de neurone :** pour faire une classification (bénigne ou maligne).
5. **Zone pour afficher les résultats :** dans cette zone, on a affiché quelques résultats comme le résultat du réseau de neurone, ainsi que la table des statistiques, et autres résultats.

4.6 Conclusion

Au cours de ce chapitre, nous nous sommes basé sur l'architecture générale de notre travail, nous avons commencé par donner une vue générale de ses principaux composants. L'approche sur laquelle se base notre étude, afin de classifier les masses mammographiques, ensuite nous avons présenté l'interface de l'application ainsi que les résultats obtenus.

Conclusion Générale

L'objectif que nous avons fixé, au début de ce mémoire, était de proposer une approche pour la sélection de caractéristiques en se basant sur les algorithmes évolutionnistes présenté dans le troisième chapitre ;

Pour la modélisation d'un système d'apprentissage automatique pour la classification d'anomalies mammaires, il est souhaitable d'utiliser un nombre optimal de caractéristiques ; en effet, il devient plus difficile de définir des limites de décision précises dans un espace de représentation très large ; un très grand nombre de caractéristique augmente les besoins de calcul. Cela signifie qu'un sous-ensemble optimal de caractéristiques doit être sélectionné pour réaliser la phase d'apprentissage automatique.

La sélection des caractéristiques pertinentes des anomalies mammaires est une étape très importante pour la réalisation d'un système de diagnostic assisté par ordinateur (DAOx) efficace. Le succès d'un système de classification dépend principalement des caractéristiques sélectionnées et de l'information fournie pour leur rôle dans le modèle. Certaines des caractéristiques extraites à partir des régions d'intérêts des images mammographiques ne sont pas significatives lorsqu'elles sont utilisées seules dans un système DAOx, mais en combinaison avec d'autres caractéristiques elles peuvent être significatives pour la classification. En général, la raison de la sélection des caractéristiques est triple :

a. Amélioration de la performance de classification du système;
b. Classification plus rapide et plus rentable;
c. Meilleure compréhension des processus qui génèrent le système DAOx;

L'entrée de notre système est une base de 59 caractéristiques calculées pour 114 masses mammographiques, cette base a été créé par les étudiants de l'université de Guelma sahtel et benkirat [35].

Cette méthode consiste à utiliser la puissance exploratoire des algorithmes génétiques. Les gènes dans cette approche se sont les caractéristiques, et un chromosome est un sous-ensemble de caractéristique, l'évaluation des chromosomes se fait par un perceptron multi-couches (PMC), Cette implémentation utilise un algorithme génétique pour faire évoluer une population de caractéristiques vers un sous-ensemble qui présente un taux de

reconnaissance maximum après un nombre d'itération, ceci est considéré comme le sous-ensemble approximatif des descripteurs les plus informatifs.

Notre système se compose principalement de 4 étapes, en premier lieu une population initiale de *N* chromosomes a été générée aléatoirement, Ensuite chaque individu est évalué par la fonction appropriée au problème (fonction de fitness),après cette phase nous avons créé une nouvelle population de *N* chromosomes par l'application d'opérateurs génétiques (croisement et mutation).L'opération se répète et faire le retour tant que la condition d'arrêt du problème n'est pas satisfaite.

Les résultats de la sélection des données obtenues de notre approche sont encourageants, il a été observé qu'un sous-ensemble de 40 caractéristiques atteint un meilleur taux de reconnaissance (88%), et les caractéristiques les plus informatifs sont présentées au quatrième chapitre.

Perspective

Le travail que nous avons proposé est une approche qui reste en développement, plusieurs améliorations peuvent lui être apportées. Il est souhaitable d'utiliser d'autres techniques au niveau des opérateurs génétiques (Sélection, croisement et mutation). Dans les travaux futurs nous devons utiliser d'autres méthode de classification afin d'améliorer le taux de reconnaissance.

Annexe

1- Corrélation

$$C1 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}\left[\frac{(i-\mu_i^d)(j-\mu_j^d)}{\sigma_i^d \sigma_j^d}\right] \quad (1)$$

2- Homogénéité

$$C2 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\left[\frac{M_{i,j}^{d}}{1+(i-j)^2}\right] \quad (2)$$

3- Variance

$$C3 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}(i-\mu^d)^2 \quad (3)$$

4- Contraste

$$C4 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}(i-j)^2 \quad (4)$$

5- Énergie

$$C5 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} (M_{i,j}^{d})^2 \quad (5)$$

6- Probabilité maximale

$$C6 = \max_{i,j} M_{i,j}^{d} \quad (6)$$

7- Somme moyenne

$$C7 = \frac{1}{2}\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}(i+j) \quad (7)$$

8- Cluster proéminence

$$C8 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}(i - \mu_i^d + j - \mu_j^d)^4 \qquad 8$$

9- Écart absolu moyen

$$C9 = \frac{1}{l/2}\sum_{d=1}^{l/2}\left|x_n^d - x_{n+8}\right| \qquad 9$$

10- Minimum

$$C10 = Min_{d=1}^{l/2} x_n^d \qquad 10$$

11- Maximum

$$C11 = Max_{d=1}^{l/2} x_n^d \qquad 11$$

12- Variance

$$C12 = \frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - x_{n+8}\right)^2 \qquad 12$$

13- Skewness

$$C13 = \frac{\frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - x_{n+8}\right)^3}{\left(\frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - x_{n+8}\right)^2\right)^{3/2}} \qquad 13$$

14- Large différence emphasis (LDE)

$$C14 = \sum_{gdif=0}^{G-1} DGD(gdif).\ln(2/gdif) \qquad 14$$

15- Sharpness

$$C15 = \sum_{gdif=0}^{G-1} DGD(gdif).(gdif)^3 \qquad 15$$

16- Second Moment de DGD (SMG)

$$C16 = \sum_{gdif=0}^{G-1} (DGD(gdif))^2 \qquad 16$$

17- Long distance emphasis for large différence (LDEL)

$$C17 = \sum_{gdif=0}^{G-1} DAD(gdif).(gdif)^2 \qquad 17$$

18- Gray Level Co-occurrence Matrix (GLCM)

$$C18 = N[i, j] = \frac{P[i; j]}{\sum_i \sum_j P[i : j]} \qquad 18$$

19- La circularité

$$C19 = \frac{aire(R \cap C_{QE})}{aire(R)} \qquad 19$$

20- La rectangularité

$$C20 = \frac{aire(R)}{aire(B_E)} \qquad 20$$

21- La compacité

$$C21 = \frac{P_2}{A} \qquad 21$$

22- Mean variations

$$C22 = \frac{1}{l/2} \sum_{d=1}^{l/2} x_n^d \qquad 22$$

23- Variance variations

$$C23 = \frac{1}{l/2} \sum_{d=1}^{l/2} \left(x_n^d - x_{n+8}\right)^2 \qquad 23$$

24- Skewness variations

$$C24 = \frac{\frac{1}{l/2} \sum_{d=1}^{l/2} \left(x_n^d - x_{n+8}\right)^3}{\left(\frac{1}{l/2} \sum_{d=1}^{l/2} (x_n^d - x_{n+8})^2\right)^{3/2}} \qquad 24$$

25- Kurtosis variations

$$C25 = \frac{\left(1/D\right) \sum_{j=1}^{D} \left(P_i^j - \overline{P}_i\right)^4}{\left(\left(1/d\right) \sum_{j=1}^{D} \left(P_i^j - \overline{P}_i\right)^2\right)^2} \qquad 25$$

26- Entropie variations

$$C26 = -\sum_{j=1}^{D} P_i^j \log_2 P_i^j \qquad 26$$

27- La courbure

$$C27 = \frac{1}{R} \tag{27}$$

$$R = \frac{a.b.c}{\sqrt{(a+b+c)(a-b+c)(a+b-c)(b-a+c)}}$$

28- La longueur radiale normalisée (LRN)

$$C28 = \frac{\sqrt{(x(i)-x_g)^2 + (y(i)-y_g)^2}}{\max(d(i))} \tag{28}$$

29- La moyenne de LRN (davg)

$$C29 = \frac{1}{N}\sum_{i=1}^{N} d(i) \tag{29}$$

30- La déviation standard de LRN

$$C30 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d(i)-d_{avg})^2} \tag{30}$$

31- L'Entropie (E)

$$C31 = \sum_{k=1}^{100} p_k \log(p_k) \tag{31}$$

32- La rugosité

$$C32 = \frac{1}{N}\sum_{i=1}^{N}(d(i)-d(i+1)) \tag{32}$$

33- Le rapport de surface

$$C33 = \frac{1}{d_{avg}.N}\sum_{i=1}^{N}\left(d(i) - d_{avg}\right) \quad 33$$

34- Différence des déviations standard

$$C34 = \left|\sigma - \sigma_{ma}\right| \quad 34$$

35- La différence de l'entropie

$$C35 = \sum_{k=1}^{100} p_k \log\left(p_k\right) \quad 35$$

36- Le rapport de surface modifiée

$$C6 = \frac{1}{d_{avg}.N}\sum_{i=1}^{N}\left(d(i) - d_{ma(i)}\right) \quad 36$$

37- Moyenne de kurtosis

$$C37 = \frac{(1/D)\sum_{j=1}^{D}\left(P_i^j - \overline{P}_i\right)^4}{\left((1/D)\sum_{j=1}^{D}\left(P_i^j - \overline{P}_i\right)^2\right)^2} - 3 \quad 37$$

38- Moyenne d'entropie

$$C38 = -\sum_{j=1}^{D} P_i^j \log_2 P_i^j \quad 38$$

39- Variation totale d'index de probabilité maximale

$$C39 = \left(\arg\max P_i^j\right) - \frac{D}{2} \quad 39$$

40- Déviation standard d'entropie

$$C40 = \sqrt{\frac{1}{D}\sum_{i=1}^{N}\left(P_i - \mu\right)^2} \qquad 40$$

41- Différence des déviations standard

$$C41 = \sum_{r=1}^{l/2} M_{RDM} \qquad 41$$

42- La distribution de différence moyenne

$$C42 = \sum_{gdif}^{G-1} M_{RDM.gdif} \qquad 42$$

43- Distribution de distance moyenne

$$C43 = \sum_{r=1} l/2 M_{RDM.r} \qquad 43$$

44- la variance

$$C44 = \frac{1}{N}\sum_{p=1}^{N}\left(I(p) - Moy\right)^2 \qquad 44$$

45- Moyenne

$$C45 = \frac{1}{l/2}\sum_{d=1}^{l/2} x_n^d \qquad 45$$

46- Moyenne corrélation

$$C46 = \frac{1}{l/2}\sum_{d=1}^{l/2} C1 \qquad 46$$

47- Minimum corrélation

$$C47 = Min_{d=1}^{l/2} C1 \qquad 47$$

48- Variance corrélation

$$C48 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}(i-\mu^d)^2 C1 \qquad 48$$

49- Skewness corrélation

$$C49 = \frac{\frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - C1\right)^3}{\left(\frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - C1\right)^2\right)^{3/2}} \qquad 49$$

50- Minimum Homogénéité

$$C50 = Min_{d=1}^{l/2} C2 \qquad 50$$

51- Variance Homogénéité

$$C51 = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^{d}(i-\mu^d)^2 C2 \qquad 51$$

52- Skewness Homogénéité

$$C52 = \frac{\frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - C2\right)^3}{\left(\frac{1}{l/2}\sum_{d=1}^{l/2}\left(x_n^d - C2\right)^2\right)^{3/2}} \quad 52$$

53- Mean énergie

$$C53 = \frac{1}{l/2}\sum_{d=1}^{l/2} C5 \quad 53$$

54- Mean absolue déviation énergie

$$C54 = \frac{1}{l/2}\sum_{d=1}^{l/2}\left|C5 - C53\right| \quad 54$$

55- Maximum énergie

$$C55 = Max_{d=1}^{l/2} C5 \quad 55$$

56- variance énergie

$$C56 = \frac{1}{l/2}\sum_{d=1}^{l/2}\left(C5 - C45\right)^2 \quad 56$$

57- skewnes énergie

$$C57 = \frac{\frac{1}{l/2}\sum_{d=1}^{l/2}\left(C5 - C45\right)^3}{\left(\frac{1}{L/2}\sum_{d=1}^{l/2}\left(C5 - C45\right)^2\right)^{\frac{3}{2}}} \quad 57$$

58- Compactness

$$C58 = \frac{P^2}{4\pi A} \quad 58$$

59- La taille de la masse

$$C59 = Taillemasse \quad 59$$

Bibliographie

[1] *SONKA, Milan, HLAVAC, Vaclav, et BOYLE, Roger. Image processing, analysis, and machine vision (3éme edition). Thomson Learning. 2014. ISBN :13: 978-0-495-24428-7*

[2] *THEODORIDIS, Sergios et KOUTROUMBAS, Konstantinos. Pattern recognition (2 éme édition). Vol.4 2008.ISBN-13: 978-0126858754*

[3] *Bernard DUBUISSON, Diagnostic et reconnaissance des formes, Hermès. 1990. ISBN: 9782866012403.*

[4] *PETROU, Maria et GARCÍA SEVILLA, Pedro. Image processing: dealing with texture. 2006. ISBN: 978-0-470-02628-1*

[5] *HEUTTE Laurent. Reconnaissance de Formes Processus de RdF, vol 9, n° 1.*

[6] *ARIF, Muhammad. Fusion de données; ultime étape de reconnaissance de formes: application à l'identification et à l'authentification, Thèse de doctorat. Université de Tours, 2005.*

[7] *KessentiniYousri. Reconnaissance De Formes Classification "statistique" des formes, vol 90,2014.*

[8] *HORAUD, Raduet MONGA, Olivier. Vision par ordinateur, Hermès ,1993.*

[9] *LI, Fan, YE, Mao, et CHEN, Xudong. An extension to Rough c-means clustering based on decision-theoretic Rough Sets model. International Journal of Approximate Reasoning, 55(1), p. 116-129, 2014.*

[10] *BOUCHER, Arnaud. Recalage et analyse d'un couple d'images: application aux mammographies, Thèse de doctorat. Université René Descartes-Paris V, 2013.*

[11] *DENGLER, Joachim, BEHRENS, Sabine, et DESAGA, Johann Friedrich. Segmentation of microcalcifications in mammograms. IEEE transactions on medical imaging, 12(4), p. 634-642, 1993.*

[12] *Pizer, SM , Amburn, EP, Austin, JD, Cromartie, R., Geselowitz, A., Greer, T., Zuiderveld, K. (. L'EQUALISATION DES HISTOGRAMENTS ADAPTATIFS ET SES VARIATIONS. Vision informatique , graphique et traitement d'image , 39 (3), 355-368.1987.*

[13] *Tomoko Matsubara, Hiroshi Fujita, Takeshi Hara, Satoshi Kasai, Osamu Otsuka, Yuji Hatanaka, and Tokiko Endo. Development of a new algorithm for detection of mammographic masses.In Digital Mammography, Springer, p 139-142, 1998.*

[14] *MUDIGONDA, Naga R., RANGAYYAN, Rangaraj M., et DESAUTELS, JE Leo. La détection de masses mammaires dans les mammographies par tranchage de densité et de l'analyse de champ d'écoulement de texture. IEEE Transactions on Medical Imaging , 20(12), p. 1215-1227, 2001.*

[15] *Paragios, N., &Deriche, R. Géodésiques, régions actives et ensembles de niveaux pour supervisé texture segmentation. International Journal of Computer Vision, 46 (3), p 223-247, 2002.*

[16] *LI, Chunming, KAO, Chiu-Yen, GORE, John C., et al. Minimization of region-scalable fitting energy for image segmentation. IEEE transactions on image processing, 17(10), p. 1940-1949, 2008.*

[17] *Journet Nicholas. Introduction au traitement d'images : Reconnaissance des formes,vol 26, 2011.*

[18] *MAY-LEVIN, Françoise, CLAVEL, Françoise, MONSONEGO, Joseph, et al. Techniques et méthodes nouvelles en santé publique appliquées en oncologie. Bulletin du cancer, 88(1), p. 23-34, 2001.*

[19] *AMERICAN COLLEGE OF RADIOLOGY. BI-RADS COMMITTEE. Breast imaging reporting and data system. American College of Radiology, 1998.*

[20] *IMENE CHEIKHROUHOU Epse KACHOURI. Description et classification des masses mammaires pour le diagnostic du cancer du sein. Ph. D. Thesis. University of Evry Val d'Essone ,France, 2012.*

[21] *Lanyi, M. L'analyse morphologique des microcalcifications. Dans le cancer du sein au stade précoce Springer Berlin Heidelberg, p.113-135,1985.*

[22] *D'ORSI, C. J., MENDELSON, E. B., IKEDA, D. M., et al. Breast imaging reporting and data system: ACR BI-RADS—breast imaging atlas. Reston, VA: American College of Radiology, vol.4, 2003,*

[23] *College of Radiology américaine. ACR BI-RADS: sein de rapports d'imagerie et de systèmes de données, imagerie du sein de l' Atlas. Reston, VA: American College of Radiology, 2003.*

[24] *WOLFE, John N. Breast patterns as an index of risk for developing breast cancer. American Journal of Roentgenology, 126(6), p. 1130-1137, 1976.*

[25] *Lévy, L., Suissa, M., Bokobsa, J., Tristant, H., Chiche, J. F., Martin, B., & Teman, G. Présentation de la traduction française du BI-RADS® (Breast Imaging Reporting System and Data System). Gynécologie obstétrique & fertilité, 33(5), 338-347. 2005.*

[26] *LE GAL, Michèle, CHAVANNE, Guy, et PELLIER, D. Valeur diagnostique des microcalcifications groupées découvertes par mammographies: à propos de 227 cas avec vérification histologique et sans tumeur du sein palpable. Bulletin du cancer, 71(1), p. 57-64,1984.*

[27] *E.J. Trudel, I. Trop : « Mammographie ». Titre de renseignement par l'Association Canadienne des Radiologistes (CAR), 2004*

[28] *P. Thiesse, N. Guérin : « Comprendre l'échographie mammaire ». Guides d'information et de dialogue pour différents cancers issue du projet SOR SAVOIR PATIENT. Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC) Paris, 2003.*

[29] *PALMA, Giovanni. Détection automatique des opacités en tomosynthèse numérique du sein.Thèse de doctorat. Télécom Paris Tech.2010.*

[30] *Holland, Adaptation in Natural and Artificial Systems. University of Michigan Press: Ann Arbor, 1975.*

[31] *Kennedy, J., &Eberhart, R. C. particle swarm algorithm In proceedings of the IEEE International Conference on neural networks,piscataway ,new jersey ,USA Vol. 5, p.1942,1995.*

[32] *HAUPT, Randy L. et HAUPT, Sue Ellen. Practical genetic algorithms.John Wiley & Sons, 2004.*

[33] *Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 201. Int J Cancer. 136(5), 359–86, 2015.*

[34] *Uhrhammer, N., Abdelouahab, A., Lafarge, L., Feillel, V., Dib, AB et Bignon, YJ. Mutations BRCA1 chez les patients algériens atteints de cancer du sein: fréquence élevée chez les cas jeunes et sporadiques. International Journal of Medical Sciences, 5(4), 197,2008.*

[35] *Sahtal okba, benkirat adel, sélection de caractéristiques pour la classification des masses mammographiques, mémoire de master, université Guelma 8 mai 1945,2016.*