

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 : Sciences et technologies de l'information et de la communication

Thème :

Analyse des paramètres de Haralick pour la classification des anomalies mammaires

Encadré par :

Dr. FERKOUS Chokri

Présenté par :

BENHAMAHOUM Abdelfeteh

Octobre 2020

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَالْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ

Remerciements

Je tiens à remercier en premier lieu mon encadreur **Dr.Ferkous Chokri** qui m'a fait confiance et acceptait de m'encadrer, ainsi que pour son patience, ces orientations et ces directives que sans lesquelles ce travail n'aurait pas pu être mené à bien.

Je tiens à remercier aussi tous les enseignants du département informatique qu'avec leurs compétences m'ont soutenu dans la poursuite de mes études.

Enfin je remercie tous ceux qui de près ou de loin m'ont porté de l'aide dans la réalisation de ce modeste travail.

Résumé

Le travail présenté dans ce mémoire concerne la réalisation d'un system d'aide au diagnostic CADx pour la classification des anomalies mammaires en se basant sur les paramètres d'Haralick. Le but principal est d'analyser le paramètre « d » utilisé pour le calcul de la matrice de cooccurrences GLCM, cette analyse nous permettra de déterminer la meilleure distance « d » à utiliser pour la classification des anomalies mammaires selon la terminologie BI-RADS. Nous avons constaté que la distance $d=23$ donne la meilleure précision (81.33%), mais la distance $d=8$ offre la plus grande valeur pour la mesure F1-score. Sachant la mesure F1-score est une combinaison de la précision de la sensibilité, donc elle est plus significative que la précision seule.

Abstract

The work presented in this memory concerns the realization of a CADx diagnostic aid system for the classification of breast abnormalities based on Haralick parameters. The main goal is to analyze the parameter "d" used for the calculation of the GLCM co-occurrence matrix, this analysis will allow us to determine the best distance "d" to use for the classification of breast abnormalities according to the BI-RADS terminology. We have found that the distance $d = 23$ gives the best accuracy (81.33%), but the distance $d = 8$ provides the greatest value for the F1-score measurement. Knowing the F1-score measurement is a combination of accuracy and sensitivity, so it is more meaningful than accuracy alone.

ملخص

يتعلق العمل المقدم في هذه المذكرة بإنشاء نظام مساعدة التشخيص CADx لتصنيف تشوهات الثدي بناءً على معايير Haralick الهدف الرئيسي هو تحليل المعامل "d" المستخدم لحساب مصفوفة التواجد المشترك لـ GLCM ، سيسمح لنا هذا التحليل بتحديد أفضل مسافة "d" لاستخدامها في تصنيف تشوهات الثدي وفقاً لمصطلحات BI-RADS. لقد وجدنا أن المسافة $d = 23$ تعطي أفضل دقة (81.33%) ، لكن المسافة $d = 8$ توفر أكبر قيمة لقياس F1-score. مع العلم أن قياس F1-score هو مزيج من الدقة والحساسية ، لذلك فهو أكثر أهمية من الدقة وحدها.

Table des matières

Introduction Générale	1
Chapitre I : Méthodes d'analyse De Texture d'image	
1. Introduction	4
2. Modèles de la texture	4
2.1. La Textures structurelles (macro-textures)	4
2.2. Textures aléatoires (micro-textures)	5
2.3. Textures directionnelles	5
3. Les différentes méthodes d'analyse de texture	5
3.1. Les méthodes spatio-fréquentielles	5
3.1.1. Filtres du domaine spatial	5
3.1.2. Filtres du domaine fréquentielles	8
3.1.2.1. Transformée de Fourier	8
3.1.2.2. Transformée de Gabor	9
3.1.2.3. Transformée en ondelettes discrète	10
3.1.2.4. Frames ondelettes	11
3.1.2.5. Les paquets d'ondelettes	12
3.1.2.6. Le banc de filtres de Schmid	12
3.1.2.7. Transformée de Huang-Hilbert	12
3.2. Les méthodes basées sur des modèles	13
3.2.1. Les champs de Markov	13
3.2.2. Les fractales	14
3.3. Les méthodes basées sur les statistiques	15
3.3.1. Statistique du premier ordre	15
3.3.2. Caractéristiques d'autocorrélation	16
3.3.3. Les Matrices de Cooccurrence	16
4. Conclusion	20
Chapitre II : Mammographie et dépistage du cancer du sein	
1. Introduction	22
2. Anatomie du sein	23
3. Techniques de dépistage du cancer de sein	23
3.1. La Mammographie	24
3.1.1. Déroulement de l'examen de mammographie	26
3.1.2. Limite de la mammographie	26
3.2. L'échographie	26
3.3. Imagerie par Résonance Magnétique (IRM)	28
4. Types d'anomalies détectables dans un dépistage de cancer de sein	29
4.1. Les densités asymétriques	29
4.2. Les masses	30
4.3. Les calcifications	31
4.3.1. Micro-calcifications	32
4.3.2. Macro-calcifications	32
5. Classification des anomalies mammaires	33
5.1. La classification de LeGal	33
5.2. La classification de Lanyi	34
5.3. La classification de BI-RADS	34
6. Description de différentes bases de données mammographiques	36
6.1. La base de données MIAS	36

6.2. La base de données AMDI.....	37
6.3. La base de données IRMA.....	37
6.4. La base de données DDSM.....	37
7. Conclusion.....	38

Chapitre III : Analyse des besoins et conception

1. Introduction.....	40
2. Démarche générale adoptée dans ce travail.....	40
3. La base de données DDSM.....	42
4. Extraction des images de la base et préparation de données.....	46
4.1. Description des tables de la base de données.....	46
4.1.1. La table IcsInfos.....	46
4.1.2. La table ImageInfos.....	46
4.1.3. La table anomalies.....	47
4.1.4. La table Edges.....	48
4.1.5. La table ROI.....	48
4.1.6. La table features.....	48
4.2. Implémentation de la base de données.....	48
4.3. Une application JAVA pour extraire les sous images de régions d'intérêt.....	49
4.4. Choix de la méthode de classification des anomalies en BI-RADS.....	49
5. Modèle utilisé dans ce travail.....	52
6. Conclusion.....	54

Chapitre IV : Implémentation et résultat expérimentaux

1. Introduction.....	56
2. Environnement de développement.....	56
2.1. Environnement pour la préparation des données.....	56
2.1.1. Téléchargement, Décompression, et Normalisation des images.....	57
2.1.2. Extraction et organisation des informations.....	57
2.2. Outils d'implémentation du réseau de neurones.....	59
2.2.1. Dispositifs matériels.....	59
2.2.2. Outils logiciels.....	59
2.3. Les catégories BI-RADS utilisées.....	62
2.4. Les paramètres du classifieur utilisés.....	62
3. Expérimentations et résultats.....	63
3.1. Critères d'évaluation du modèle.....	63
3.1.1. Matrice de confusion.....	63
3.1.2. Précision de classification (Accuracy).....	64
3.1.3. Le rappel ou sensibilité (Recall).....	64
3.1.4. Le F-score.....	65
3.2. Résultats des expériences.....	65
4. Conclusion.....	67

Conclusion Générale.....	68
---------------------------------	-----------

Bibliographie.....	69
---------------------------	-----------

Annexes.....	74
---------------------	-----------

Annexe 01 : Le script python « convert.py ».....	75
---	-----------

Annexe 02 : Le script Python « ljpeg.py».....	76
--	-----------

Annexe 03 : Un script Python pour calculer les paramètres de Haralick.....	77
---	-----------

Annexe 04 : Le script Python pour modéliser le réseau de neurones artificiel.....	78
--	-----------

Liste des Figures

Figure I.1 : Exemples de Textures structurales (macro-textures)	4
Figure I.2 : Exemples de Textures aléatoire (micro-textures)	5
Figure I.3 : Partition du plan fréquentiel	8
Figure I.4 : Décomposition en ondelette d'une image 2D	11
Figure I.5 : Principe de la transformée de Huang – Hilbert.	13
Figure I.6 : Construction de la courbe de Koch.	14
Figure I.7 : Formation de la matrice de cooccurrence GLCM.....	17
Figure II.1 : Carte mondiale présentant le classement national du cancer comme cause de décès pour les âges inférieurs à 70 ans en 2015	22
Figure II.2 : Schéma anatomique du sein	23
Figure II.3 Les composants d'un mammographe.	25
Figure II.4 : A : Résultat de l'incidence de face, B : Résultat de l'incidence oblique.....	25
Figure II.5 : Echographie du sein pour diagnostic du cancer.	27
Figure II.6 : Echographie du sein pour guidage lors d'une opération de biopsie.....	27
Figure II.7 : IRM mammaires coupe axiale.	29
Figure II.8 : une asymétrie de densité observée seulement sur la mammographie gauche.	30
Figure II.9 : Les différentes formes possibles d'une masse	30
Figure II.10 : Les différents contours possibles d'une masse.....	31
Figure II.11 : A- Images de distorsion architecturale. B-Images de masses avec des limites en spicules.	31
Figure II.12 : Images de distorsion architecturale.	31
Figure II.12 : Type de calcification.	33
Figure III.1 : Schéma de la démarche générale adoptée pour analyser l'utilisation des paramètres de Haralick dans la classification des anomalies mammographiques.	41
Figure III.2 : quelques échantillons de la base DDSM.....	42
Figure III.3 : La liste des fichiers du cas C-0029 volume : Cancer N° 1.	43
Figure III.4 : Exemple d'un fichier « .ics ».	44
Figure III.5 : Exemple d'un fichier « .OVERLAY».	45
Figure III.6 : Codification de déplacement dans la chaîne des contours d'un fichier « .OVERLAY».	45
Figure III.7 : Exemple d'un fichier ".16_PGM" représentant les quatre mammographies sous-échantillonnées.....	46
Figure III.8 : Model conceptuel de données pour la base DDSM.	47
Figure III.9 : Résultats de comparaison entre les 6 méthodes de classification.	51
Figure III.10 : Résultats de classification par un réseau de neurones artificiel.	52
Figure III.11 ; Modèle du réseau de neurone utilisé pour la classification des anomalies mammaires.	54
Figure IV.1 : Les couches de la bibliothèque TensorFlow	60
Figure IV.2 Exemple de résultat de l'entraînement du modèle	66

Liste des Tableaux

Tableau I.1 : Les masques de Laws 3x3	6
Tableau I.2 : Les masques de Laws 5x5	7
Tableau I.3 : Les 14 paramètres de Haralick.	19
Tableau II.1 : Classification de Le Gal.....	33
Tableau II.2 : Classification BIRADS de l'ACR.....	35
Tableau II.3 : Conduite à tenir pour chaque classe de l'ACR.....	36
Tableau III.1 : Les différents sites et scanners utilisé dans la numérisation de la base DDSM.....	43
Tableau III.2 : Répartition des cas selon le type de pathologie Dans la base DDSM.	43
Tableau IV.1 : Description des couches de la bibliothèque TensorFlow	60
Tableau IV.2 : Matrice de confusion.	64
Tableau IV.3 Résultats des 20 distances.....	68

Liste des Abréviations

GLCM	Grey Level Co-occurrence Matrix
OMS	Organisation mondial de la santé
WHO	World Health Organization
DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
ACR	American College of Radiology
CADx	Computed Assisted Diagnosis
BI-RADS	SYSTEME DE COMPTES-RENDUS ET DE DONNEES D'IMAGERIE MAMMAIRE (BREAST IMAGING REPORTING AND DATA SYSTEM)
CC	CRANIO CAUDAL
IRM	L'Imagerie par Résonance Magnétique
FN	Faux Négatifs
LJPEG	LOSSLESS JPEG
DDSM	LA BASE DE DONNEES NUMERIQUE POUR LA MAMMOGRAPHIE DE DEPISTAGE
MIAS	Mammographic Image Analysis Society
RNA	Réseau de Neurone Artificiel
MCC	MICRO CALCIFICATION
MLO	MEDIO-LATERALE OBLIQUE
PMC	PERCEPTRON MULTICOUCHE
ROIs	REGION D'INTERET
JPA	Java Persistence Api
ORM	Object Relational Mapping
IDE	Integrated Development Environment
JVM	Java Virtual Machine
LDA	Linear Discriminant Analysis
KNN	Méthode des k plus proches voisins
RL	Régressions Logique
DTC	Decision Tree Classifier
NB	Naïve Bayes

Introduction Générale

Le cancer du sein est le plus fréquent des cancers féminins à travers le monde, il signifie la présence de cellules anormales qui se multiplient de façon incontrôlée. Selon les estimations de l'OMS [5], les maladies mammaires touchent 2.1 millions de femmes chaque année, il y a eu 627 000 décès au niveau mondial en 2018. Le dépistage précoce du cancer augmente les chances de guérison et permet dans plusieurs cas un traitement plus efficace.

La plus importante étape lors du dépistage est sans doute la détection des anomalies sur les mammographies et leurs classifications selon le degré de soupçon accordé à la présence ou non d'un cancer afin de prendre les mesures adéquates, Car la détection précoce du cancer est le facteur principal qui agit sur la guérison. Cette tâche est réalisée habituellement par un radiologue spécialisé dans le domaine.

Pour aider les radiologues à l'accomplissement de cette tâche de façon automatique, nous voulons proposer dans ce travail un système d'aide au diagnostic CADx pour la classification des anomalies mammaires en se basant sur les méthodes statistiques du second ordre. Ces méthodes mis en jeu la relation de voisinage des pixels dans la région d'intérêt (ROI).

La méthode du second ordre la plus utilisée pour le traitement des images médicale notamment les mammographies est celle introduite par Haralick [34] en 1973. Cette méthode se base sur le calcul d'une matrice de cooccurrence dit GLCM (Grey Level Co-occurrence Matrix), comptabilisant les nombres de fois que deux niveaux de gris apparaissent dans l'image à un distance et un angle donné, et qui nous permettra par la suite de calculer les 14 caractéristiques de Haralick.

Le but principale de notre travail est l'analyse de l'influence de la distance « d » utilisé pour le calcul de la matrice GLCM sur le résultat de classification des anomalies détectées dans les mammographies.

Notre travail se divise donc en deux principales parties :

1. la première partie concerne la préparation des données d'apprentissage, elle consiste à réaliser les tâches suivantes :

- Les prétraitements de décodage des mammographies et de la base DDSM.
- L'extraction des régions d'intérêts des anomalies en se basant sur les annotations de la base de données.
- Calculer les caractéristiques d'Haralick pour créer la base d'apprentissage et de test (une pour chaque distance « d »).

2. **La deuxième partie** concerne la réalisation des expériences et des tests afin de déterminer la meilleure distance, cette partie se compose des tâches suivantes :

- Construire a réseau de neurones artificiel pour tester les 20 ensembles d'apprentissage.
- Comparer les résultats et déduire la meilleure distance à utiliser dans le calcul des matrices de cooccurrences pour la classification des anomalies mammaires.

Pour la réalisation de la réalisation la première partie nous allons concevoir une application JAVA pour la gestion des informations sur la base de données mammographique DDSM, Cette approche nous a été imposée par la complexité de ces informations. Grace à cette application nous auront un accès flexible aux images et aux informations sur les anomalies de la base de données, ce qui nous facilite l'extraction des caractéristiques et la création des ensembles d'apprentissages.

La deuxième partie sera réalisées en utilisant le langage Python sous l'environnement colab de Google, à l'aide d'un réseau de neurone conçu en utilisant la bibliothèque Keras et TensorFow.

CHAPITRE I

METHODES
D'ANALYSE DE
TEXTURE D'IMAGE

I.1 Introduction

La texture est reconnaissable à la fois tactile et optique, la texture tactile fait référence à la sensation tangible d'une surface. La texture optique (visuelle) fait référence à la forme ou au contenu de l'image [1]. Dans un système de vision humaine le diagnostic de la texture est facilement réalisable, mais dans le domaine de la vision par ordinateur et du traitement d'images numériques ont leurs propres complexités lors de l'analyse texturale.

En traitement d'image, la texture peut être définie en fonction de la variation spatiale de l'intensité lumineuse des pixels. Elle est le terme principal utilisé pour définir des objets ou des concepts dans une image. L'analyse texturale joue un rôle important dans plusieurs sous-domaines de la vision par ordinateur tels que la reconnaissance d'objets, détection de défauts de surface, reconnaissance de formes, analyse d'images médicales, etc. [2]

Dans ce premier chapitre nous présentons un état de l'art sur la textures de l'image, et voir les principales méthodes et approches utilisés lors de leur l'analyse.

I.2 Modèles de la texture

La notion de texture est tellement complexe ce qui impose une grande difficulté lors de la classification de l'ensemble des textures. En effet, ils existent une multitude de textures qu'on peut regrouper en trois principales familles [4] :

1. Textures structurelles.
2. Textures aléatoires.
3. Textures directionnelles.

I.2.1 La Textures structurelles (macro-textures)

Cette texture se caractérise principalement par la répétition d'un motif de base appelé texton suivant des règles de directions et de placement. On cite comme exemple de cette texture, la texture de mur de briques, la texture de certains tissus ou d'un grillage (Fig. 1.1).



Figure I.1 : Exemples de Textures structurelles (macro-textures)

I.2.2 Textures aléatoires (micro-textures)

Dans cette texture aucun motif régulier n'est détectable ou localisable. Cependant, leur impression visuelle reste globalement homogène et représente des textures microscopiques. Comme exemple de cette texture on trouve les surfaces d'eau, de bois et de différents matériaux, les champs de pâturage et d'herbe.



Figure I.2 : Exemples de Textures aléatoire (micro-textures)

I.2.3 Textures directionnelles

Ces textures n'ont pas un motif structurant de base, Néanmoins, ne sont pas totalement aléatoire, elles se caractérisent par une structure orienté.

I.3 Les différentes méthodes d'analyse de texture

De nombreuses méthodes ont été proposées pour extraire et manipuler la texture des images numériques [1]. Les méthodes d'analyse de texture sont généralement classées en quatre catégories : méthodes statistiques, méthodes structurelles, méthodes basées sur des modèles et les méthodes basées sur des transformations fréquentielles.

I.3.1 Les méthodes spatio-fréquentielles :

La recherche psychophysique a montré que le cerveau humain effectue une analyse fréquentielle de l'image [3] [4]. La texture est particulièrement adaptée à ce type d'analyse en raison de ses propriétés.

Dans ce type de méthodes, les propriétés de la texture sont extraites à partir des caractéristiques fréquentielles de l'image. Certaines de ces méthodes sont basées sur le filtrage spatial (domaine spatial), d'autres extraient les attributs de texture directement à partir de la représentation de l'image dans le domaine de Fourier et d'autres exploitent à la fois le domaine spatial et le domaine fréquentiel lors du calcul des attributs.

I.3.1.1. Filtres du domaine spatial

Les filtres de domaine spatial sont le moyen le plus direct pour capturer les propriétés de la texture d'une image. Les principes de définition de ces méthodes sont concentrés sur la mesure de la densité des contours par unité de surface. Les textures fines ont tendance à avoir une densité de contours par unité de surface plus élevée que les textures plus gros.

Le filtrage consiste à l'application d'une convolution linéaire sur l'image par un noyau de petite taille (3×3 ou 5×5). Les masques de Laws [6] [7] est l'exemple le plus connu de ces filtres. Les masques de Laws 3x3 sont obtenus, par combinaison des trois masques 1 × 3 de base suivants :

$$L3 = [1 \ 2 \ 1]$$

$$E3 = [-1 \ 0 \ 1]$$

$$S3 = [-1 \ 2 \ -1]$$

Les lettres initiales de ces masques indiquent respectivement : moyenne locale de niveaux de gris (Local averaging), détection de contours (Edge detection) et la détection de points (Spot detection). De ces trois masque de base de dimension 1x3 on combine neuf masque de taille 3×3 qui représente le masque de Laws 3×3 (Tableau I.1.).

L3L3	E3L3	S3L3
1 2 1	-1 0 1	-1 2 -1
2 4 2	-2 0 2	-2 4 -2
1 2 1	-1 0 1	-1 2 -1
E3L3	E3E3	E3S3
-1 -2 -1	1 0 -1	1 -2 1
0 0 0	0 0 0	0 0 0
1 2 1	-1 0 1	-1 2 -1
S3L3	S3E3	S3S3
-1 -2 -1	1 0 -1	1 -2 1
2 4 2	-2 0 2	-2 4 -2
-1 -2 -1	1 0 -1	1 -2 1

Tableau I.1 : Les masques de Laws 3x3

Pour les masques de taille (5×5) nous pouvons utiliser les cinq masques de base (1×5) suivant :

$$L5 = [1 \ 4 \ 6 \ 4 \ 1] ;$$

$$E5 = [-1 \ -2 \ 0 \ 2 \ 1] ;$$

$$S5 = [-1 \ 0 \ 2 \ 0 \ -1] ;$$

$$R5 = [1 \ -4 \ 6 \ -4 \ 1] ;$$

$$W5 = [-1 \ 20 \ -2 \ 1] ;$$

Les initiales des masques L, E et S sont comme dans les masques 1×3, le masque R5 représente la détection des ondulations (Ripes detection) et le masque W5 représente la détection des ondes (Wavre detection).

La combinaison de ces cinq masques nous donne 25 masques de taille (5×5) (Tableau I.2)

<p style="text-align: center;">L5L5</p> <table border="1"> <tr><td>1</td><td>4</td><td>6</td><td>4</td><td>1</td></tr> <tr><td>4</td><td>16</td><td>24</td><td>16</td><td>4</td></tr> <tr><td>6</td><td>24</td><td>36</td><td>24</td><td>6</td></tr> <tr><td>4</td><td>16</td><td>24</td><td>16</td><td>4</td></tr> <tr><td>1</td><td>4</td><td>6</td><td>4</td><td>1</td></tr> </table>	1	4	6	4	1	4	16	24	16	4	6	24	36	24	6	4	16	24	16	4	1	4	6	4	1	<p style="text-align: center;">L5E5</p> <table border="1"> <tr><td>-1</td><td>-4</td><td>-6</td><td>-4</td><td>-1</td></tr> <tr><td>-2</td><td>-8</td><td>-12</td><td>-8</td><td>-2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>8</td><td>12</td><td>8</td><td>2</td></tr> <tr><td>1</td><td>4</td><td>6</td><td>4</td><td>1</td></tr> </table>	-1	-4	-6	-4	-1	-2	-8	-12	-8	-2	0	0	0	0	0	2	8	12	8	2	1	4	6	4	1	<p style="text-align: center;">L5S5</p> <table border="1"> <tr><td>-1</td><td>-4</td><td>-6</td><td>-4</td><td>-1</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>8</td><td>12</td><td>8</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-1</td><td>-4</td><td>-6</td><td>-4</td><td>-1</td></tr> </table>	-1	-4	-6	-4	-1	0	0	0	0	0	2	8	12	8	2	0	0	0	0	0	-1	-4	-6	-4	-1	<p style="text-align: center;">L5W5</p> <table border="1"> <tr><td>-1</td><td>-4</td><td>-6</td><td>-4</td><td>-1</td></tr> <tr><td>2</td><td>8</td><td>12</td><td>8</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>-8</td><td>-12</td><td>-8</td><td>-2</td></tr> <tr><td>1</td><td>4</td><td>6</td><td>4</td><td>1</td></tr> </table>	-1	-4	-6	-4	-1	2	8	12	8	2	0	0	0	0	0	-2	-8	-12	-8	-2	1	4	6	4	1	<p style="text-align: center;">L5R5</p> <table border="1"> <tr><td>1</td><td>4</td><td>6</td><td>4</td><td>1</td></tr> <tr><td>-4</td><td>-16</td><td>-24</td><td>-16</td><td>-4</td></tr> <tr><td>6</td><td>24</td><td>36</td><td>24</td><td>6</td></tr> <tr><td>-4</td><td>-16</td><td>-24</td><td>-16</td><td>-4</td></tr> <tr><td>1</td><td>4</td><td>6</td><td>4</td><td>1</td></tr> </table>	1	4	6	4	1	-4	-16	-24	-16	-4	6	24	36	24	6	-4	-16	-24	-16	-4	1	4	6	4	1
1	4	6	4	1																																																																																																																													
4	16	24	16	4																																																																																																																													
6	24	36	24	6																																																																																																																													
4	16	24	16	4																																																																																																																													
1	4	6	4	1																																																																																																																													
-1	-4	-6	-4	-1																																																																																																																													
-2	-8	-12	-8	-2																																																																																																																													
0	0	0	0	0																																																																																																																													
2	8	12	8	2																																																																																																																													
1	4	6	4	1																																																																																																																													
-1	-4	-6	-4	-1																																																																																																																													
0	0	0	0	0																																																																																																																													
2	8	12	8	2																																																																																																																													
0	0	0	0	0																																																																																																																													
-1	-4	-6	-4	-1																																																																																																																													
-1	-4	-6	-4	-1																																																																																																																													
2	8	12	8	2																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	-8	-12	-8	-2																																																																																																																													
1	4	6	4	1																																																																																																																													
1	4	6	4	1																																																																																																																													
-4	-16	-24	-16	-4																																																																																																																													
6	24	36	24	6																																																																																																																													
-4	-16	-24	-16	-4																																																																																																																													
1	4	6	4	1																																																																																																																													
<p style="text-align: center;">E5L5</p> <table border="1"> <tr><td>-1</td><td>-2</td><td>0</td><td>2</td><td>1</td></tr> <tr><td>-4</td><td>-8</td><td>0</td><td>8</td><td>4</td></tr> <tr><td>-6</td><td>-12</td><td>0</td><td>12</td><td>6</td></tr> <tr><td>-4</td><td>-8</td><td>0</td><td>8</td><td>4</td></tr> <tr><td>-1</td><td>-2</td><td>0</td><td>2</td><td>1</td></tr> </table>	-1	-2	0	2	1	-4	-8	0	8	4	-6	-12	0	12	6	-4	-8	0	8	4	-1	-2	0	2	1	<p style="text-align: center;">E5E5</p> <table border="1"> <tr><td>1</td><td>2</td><td>0</td><td>-2</td><td>-1</td></tr> <tr><td>2</td><td>4</td><td>0</td><td>-4</td><td>-2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>-4</td><td>0</td><td>4</td><td>2</td></tr> <tr><td>-1</td><td>-2</td><td>0</td><td>2</td><td>1</td></tr> </table>	1	2	0	-2	-1	2	4	0	-4	-2	0	0	0	0	0	-2	-4	0	4	2	-1	-2	0	2	1	<p style="text-align: center;">E5S5</p> <table border="1"> <tr><td>1</td><td>2</td><td>0</td><td>-2</td><td>-1</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>-4</td><td>0</td><td>4</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>2</td><td>0</td><td>-2</td><td>-1</td></tr> </table>	1	2	0	-2	-1	0	0	0	0	0	-2	-4	0	4	2	0	0	0	0	0	1	2	0	-2	-1	<p style="text-align: center;">E5W5</p> <table border="1"> <tr><td>1</td><td>2</td><td>0</td><td>-2</td><td>-1</td></tr> <tr><td>-2</td><td>-4</td><td>0</td><td>4</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>4</td><td>0</td><td>-4</td><td>-2</td></tr> <tr><td>-1</td><td>-2</td><td>0</td><td>2</td><td>1</td></tr> </table>	1	2	0	-2	-1	-2	-4	0	4	2	0	0	0	0	0	2	4	0	-4	-2	-1	-2	0	2	1	<p style="text-align: center;">E5R5</p> <table border="1"> <tr><td>-1</td><td>-2</td><td>0</td><td>2</td><td>1</td></tr> <tr><td>4</td><td>8</td><td>0</td><td>-8</td><td>-4</td></tr> <tr><td>-6</td><td>-12</td><td>0</td><td>12</td><td>6</td></tr> <tr><td>4</td><td>8</td><td>0</td><td>-8</td><td>-4</td></tr> <tr><td>-1</td><td>-2</td><td>0</td><td>2</td><td>1</td></tr> </table>	-1	-2	0	2	1	4	8	0	-8	-4	-6	-12	0	12	6	4	8	0	-8	-4	-1	-2	0	2	1
-1	-2	0	2	1																																																																																																																													
-4	-8	0	8	4																																																																																																																													
-6	-12	0	12	6																																																																																																																													
-4	-8	0	8	4																																																																																																																													
-1	-2	0	2	1																																																																																																																													
1	2	0	-2	-1																																																																																																																													
2	4	0	-4	-2																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	-4	0	4	2																																																																																																																													
-1	-2	0	2	1																																																																																																																													
1	2	0	-2	-1																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	-4	0	4	2																																																																																																																													
0	0	0	0	0																																																																																																																													
1	2	0	-2	-1																																																																																																																													
1	2	0	-2	-1																																																																																																																													
-2	-4	0	4	2																																																																																																																													
0	0	0	0	0																																																																																																																													
2	4	0	-4	-2																																																																																																																													
-1	-2	0	2	1																																																																																																																													
-1	-2	0	2	1																																																																																																																													
4	8	0	-8	-4																																																																																																																													
-6	-12	0	12	6																																																																																																																													
4	8	0	-8	-4																																																																																																																													
-1	-2	0	2	1																																																																																																																													
<p style="text-align: center;">S5L5</p> <table border="1"> <tr><td>-1</td><td>0</td><td>2</td><td>0</td><td>-1</td></tr> <tr><td>-4</td><td>0</td><td>8</td><td>0</td><td>-4</td></tr> <tr><td>-6</td><td>0</td><td>12</td><td>0</td><td>-6</td></tr> <tr><td>-4</td><td>0</td><td>8</td><td>0</td><td>-4</td></tr> <tr><td>-1</td><td>0</td><td>2</td><td>0</td><td>-1</td></tr> </table>	-1	0	2	0	-1	-4	0	8	0	-4	-6	0	12	0	-6	-4	0	8	0	-4	-1	0	2	0	-1	<p style="text-align: center;">S5E5</p> <table border="1"> <tr><td>1</td><td>0</td><td>-2</td><td>0</td><td>1</td></tr> <tr><td>2</td><td>0</td><td>-4</td><td>0</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>0</td><td>4</td><td>0</td><td>-2</td></tr> <tr><td>-1</td><td>0</td><td>2</td><td>0</td><td>-1</td></tr> </table>	1	0	-2	0	1	2	0	-4	0	2	0	0	0	0	0	-2	0	4	0	-2	-1	0	2	0	-1	<p style="text-align: center;">S5S5</p> <table border="1"> <tr><td>1</td><td>0</td><td>-2</td><td>0</td><td>1</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>0</td><td>4</td><td>0</td><td>-2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>-2</td><td>0</td><td>1</td></tr> </table>	1	0	-2	0	1	0	0	0	0	0	-2	0	4	0	-2	0	0	0	0	0	1	0	-2	0	1	<p style="text-align: center;">S5W5</p> <table border="1"> <tr><td>1</td><td>0</td><td>-2</td><td>0</td><td>1</td></tr> <tr><td>-2</td><td>0</td><td>4</td><td>0</td><td>-2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>0</td><td>-4</td><td>0</td><td>2</td></tr> <tr><td>-1</td><td>0</td><td>2</td><td>0</td><td>-1</td></tr> </table>	1	0	-2	0	1	-2	0	4	0	-2	0	0	0	0	0	2	0	-4	0	2	-1	0	2	0	-1	<p style="text-align: center;">S5R5</p> <table border="1"> <tr><td>-1</td><td>0</td><td>2</td><td>0</td><td>-1</td></tr> <tr><td>4</td><td>0</td><td>-8</td><td>0</td><td>4</td></tr> <tr><td>-6</td><td>0</td><td>12</td><td>0</td><td>-6</td></tr> <tr><td>4</td><td>0</td><td>-8</td><td>0</td><td>4</td></tr> <tr><td>-1</td><td>0</td><td>2</td><td>0</td><td>-1</td></tr> </table>	-1	0	2	0	-1	4	0	-8	0	4	-6	0	12	0	-6	4	0	-8	0	4	-1	0	2	0	-1
-1	0	2	0	-1																																																																																																																													
-4	0	8	0	-4																																																																																																																													
-6	0	12	0	-6																																																																																																																													
-4	0	8	0	-4																																																																																																																													
-1	0	2	0	-1																																																																																																																													
1	0	-2	0	1																																																																																																																													
2	0	-4	0	2																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	0	4	0	-2																																																																																																																													
-1	0	2	0	-1																																																																																																																													
1	0	-2	0	1																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	0	4	0	-2																																																																																																																													
0	0	0	0	0																																																																																																																													
1	0	-2	0	1																																																																																																																													
1	0	-2	0	1																																																																																																																													
-2	0	4	0	-2																																																																																																																													
0	0	0	0	0																																																																																																																													
2	0	-4	0	2																																																																																																																													
-1	0	2	0	-1																																																																																																																													
-1	0	2	0	-1																																																																																																																													
4	0	-8	0	4																																																																																																																													
-6	0	12	0	-6																																																																																																																													
4	0	-8	0	4																																																																																																																													
-1	0	2	0	-1																																																																																																																													
<p style="text-align: center;">W5L5</p> <table border="1"> <tr><td>-1</td><td>2</td><td>0</td><td>-2</td><td>1</td></tr> <tr><td>-4</td><td>8</td><td>0</td><td>-8</td><td>4</td></tr> <tr><td>-6</td><td>12</td><td>0</td><td>-12</td><td>6</td></tr> <tr><td>-4</td><td>8</td><td>0</td><td>-8</td><td>4</td></tr> <tr><td>-1</td><td>2</td><td>0</td><td>-2</td><td>1</td></tr> </table>	-1	2	0	-2	1	-4	8	0	-8	4	-6	12	0	-12	6	-4	8	0	-8	4	-1	2	0	-2	1	<p style="text-align: center;">W5E5</p> <table border="1"> <tr><td>1</td><td>-2</td><td>0</td><td>2</td><td>-1</td></tr> <tr><td>2</td><td>-4</td><td>0</td><td>4</td><td>-2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>4</td><td>0</td><td>-4</td><td>2</td></tr> <tr><td>-1</td><td>2</td><td>0</td><td>-2</td><td>1</td></tr> </table>	1	-2	0	2	-1	2	-4	0	4	-2	0	0	0	0	0	-2	4	0	-4	2	-1	2	0	-2	1	<p style="text-align: center;">W5S5</p> <table border="1"> <tr><td>1</td><td>-2</td><td>0</td><td>2</td><td>-1</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>4</td><td>0</td><td>-4</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>-2</td><td>0</td><td>2</td><td>-1</td></tr> </table>	1	-2	0	2	-1	0	0	0	0	0	-2	4	0	-4	2	0	0	0	0	0	1	-2	0	2	-1	<p style="text-align: center;">W5W5</p> <table border="1"> <tr><td>1</td><td>-2</td><td>0</td><td>2</td><td>-1</td></tr> <tr><td>-2</td><td>4</td><td>0</td><td>-4</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>-4</td><td>0</td><td>4</td><td>-2</td></tr> <tr><td>-1</td><td>2</td><td>0</td><td>-2</td><td>1</td></tr> </table>	1	-2	0	2	-1	-2	4	0	-4	2	0	0	0	0	0	2	-4	0	4	-2	-1	2	0	-2	1	<p style="text-align: center;">W5R5</p> <table border="1"> <tr><td>1</td><td>-2</td><td>0</td><td>2</td><td>-1</td></tr> <tr><td>4</td><td>8</td><td>0</td><td>-8</td><td>-4</td></tr> <tr><td>-6</td><td>12</td><td>0</td><td>-12</td><td>6</td></tr> <tr><td>4</td><td>-8</td><td>0</td><td>8</td><td>-4</td></tr> <tr><td>-1</td><td>2</td><td>0</td><td>-2</td><td>1</td></tr> </table>	1	-2	0	2	-1	4	8	0	-8	-4	-6	12	0	-12	6	4	-8	0	8	-4	-1	2	0	-2	1
-1	2	0	-2	1																																																																																																																													
-4	8	0	-8	4																																																																																																																													
-6	12	0	-12	6																																																																																																																													
-4	8	0	-8	4																																																																																																																													
-1	2	0	-2	1																																																																																																																													
1	-2	0	2	-1																																																																																																																													
2	-4	0	4	-2																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	4	0	-4	2																																																																																																																													
-1	2	0	-2	1																																																																																																																													
1	-2	0	2	-1																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	4	0	-4	2																																																																																																																													
0	0	0	0	0																																																																																																																													
1	-2	0	2	-1																																																																																																																													
1	-2	0	2	-1																																																																																																																													
-2	4	0	-4	2																																																																																																																													
0	0	0	0	0																																																																																																																													
2	-4	0	4	-2																																																																																																																													
-1	2	0	-2	1																																																																																																																													
1	-2	0	2	-1																																																																																																																													
4	8	0	-8	-4																																																																																																																													
-6	12	0	-12	6																																																																																																																													
4	-8	0	8	-4																																																																																																																													
-1	2	0	-2	1																																																																																																																													
<p style="text-align: center;">R5L5</p> <table border="1"> <tr><td>1</td><td>-4</td><td>6</td><td>-4</td><td>1</td></tr> <tr><td>4</td><td>16</td><td>24</td><td>-16</td><td>4</td></tr> <tr><td>6</td><td>-24</td><td>36</td><td>-24</td><td>6</td></tr> <tr><td>4</td><td>-16</td><td>24</td><td>-16</td><td>4</td></tr> <tr><td>1</td><td>-4</td><td>6</td><td>-4</td><td>1</td></tr> </table>	1	-4	6	-4	1	4	16	24	-16	4	6	-24	36	-24	6	4	-16	24	-16	4	1	-4	6	-4	1	<p style="text-align: center;">R5E5</p> <table border="1"> <tr><td>-1</td><td>4</td><td>-6</td><td>4</td><td>-1</td></tr> <tr><td>-2</td><td>8</td><td>12</td><td>8</td><td>-2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>-8</td><td>12</td><td>-8</td><td>2</td></tr> <tr><td>1</td><td>-4</td><td>6</td><td>-4</td><td>1</td></tr> </table>	-1	4	-6	4	-1	-2	8	12	8	-2	0	0	0	0	0	2	-8	12	-8	2	1	-4	6	-4	1	<p style="text-align: center;">R5S5</p> <table border="1"> <tr><td>-1</td><td>4</td><td>-6</td><td>4</td><td>-1</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>2</td><td>-8</td><td>12</td><td>-8</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-1</td><td>4</td><td>-6</td><td>4</td><td>-1</td></tr> </table>	-1	4	-6	4	-1	0	0	0	0	0	2	-8	12	-8	2	0	0	0	0	0	-1	4	-6	4	-1	<p style="text-align: center;">R5W5</p> <table border="1"> <tr><td>-1</td><td>4</td><td>-6</td><td>4</td><td>-1</td></tr> <tr><td>2</td><td>-8</td><td>12</td><td>-8</td><td>2</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-2</td><td>8</td><td>12</td><td>8</td><td>-2</td></tr> <tr><td>1</td><td>-4</td><td>6</td><td>-4</td><td>1</td></tr> </table>	-1	4	-6	4	-1	2	-8	12	-8	2	0	0	0	0	0	-2	8	12	8	-2	1	-4	6	-4	1	<p style="text-align: center;">R5R5</p> <table border="1"> <tr><td>1</td><td>-4</td><td>6</td><td>-4</td><td>1</td></tr> <tr><td>-4</td><td>16</td><td>-24</td><td>16</td><td>-4</td></tr> <tr><td>6</td><td>-24</td><td>36</td><td>-24</td><td>6</td></tr> <tr><td>-4</td><td>16</td><td>-24</td><td>16</td><td>-4</td></tr> <tr><td>1</td><td>-4</td><td>6</td><td>-4</td><td>1</td></tr> </table>	1	-4	6	-4	1	-4	16	-24	16	-4	6	-24	36	-24	6	-4	16	-24	16	-4	1	-4	6	-4	1
1	-4	6	-4	1																																																																																																																													
4	16	24	-16	4																																																																																																																													
6	-24	36	-24	6																																																																																																																													
4	-16	24	-16	4																																																																																																																													
1	-4	6	-4	1																																																																																																																													
-1	4	-6	4	-1																																																																																																																													
-2	8	12	8	-2																																																																																																																													
0	0	0	0	0																																																																																																																													
2	-8	12	-8	2																																																																																																																													
1	-4	6	-4	1																																																																																																																													
-1	4	-6	4	-1																																																																																																																													
0	0	0	0	0																																																																																																																													
2	-8	12	-8	2																																																																																																																													
0	0	0	0	0																																																																																																																													
-1	4	-6	4	-1																																																																																																																													
-1	4	-6	4	-1																																																																																																																													
2	-8	12	-8	2																																																																																																																													
0	0	0	0	0																																																																																																																													
-2	8	12	8	-2																																																																																																																													
1	-4	6	-4	1																																																																																																																													
1	-4	6	-4	1																																																																																																																													
-4	16	-24	16	-4																																																																																																																													
6	-24	36	-24	6																																																																																																																													
-4	16	-24	16	-4																																																																																																																													
1	-4	6	-4	1																																																																																																																													

Tableau I.2 : Les masques de Laws 5×5

Pour obtenir les attributs de texture on applique les filtres sur l'image, et pour chaque image filtrée on calcule une énergie locale pour chaque pixel à partir des pixels voisins, de cette énergie on extrait les propriétés statistiques locales (moyenne et variance) qui sont alors considérées comme des attributs de textures.

I.3.1.2. Filtres du domaine fréquentielles

La représentation fréquentielles de l'image obtenu soit par la transformée de Fourier discrète (DFT : Discrete Fourier Transform) ou la transformée de cosinus discrète (DCT : Discrete Cosine Transform) peut bien caractériser les attributs de la texture [8], les variations douces des niveaux de gris dans l'image est représentée par de basses fréquences, tandis que les haut fréquences représente le changement brutales des niveaux de gris. Pour caractériser les textures, on applique un ensemble de filtres, chacun étant sensible à une fréquence particulière. Un ou plusieurs attributs statistiques (comme l'énergie, la variance,...) sont ensuite calculés à partir des images filtrées.

I.3.1.2.1 Transformée de Fourier

La transformée de Fourier [4] [12] est une des méthodes utilisées pour l'analyse linéaire d'une image. La transformée de Fourier $F(u, v)$ d'une image I de dimension $N \times M$ prend comme valeur:

$$F(u, v) = \frac{1}{N \times M} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} I(x, y) e^{-i2\pi(u\frac{x}{N} + v\frac{y}{M})}$$

Pour extraire les caractéristiques de la texture de l'image directement à partir de sa représentation dans le domaine de Fourier, On divise le domaine des fréquences (plan u- v) soit en secteur angulaire soit en anneaux centrique et on calcule l'énergie de ces régions qui représente une caractéristique de la texture.

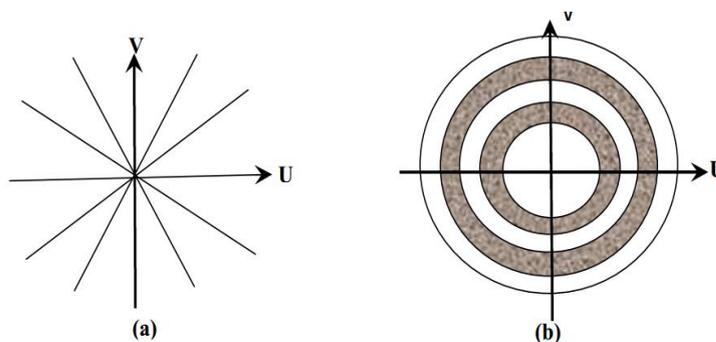


Figure I.3 : Partition du plan fréquentiel en (a): secteurs angulaires et (b): anneaux concentriques

La transformée de Fourier est un outil bien adapté à l'analyse des informations globale de l'image, cependant elle n'est pas appropriée à l'analyse de formes localisées spatialement. Pour utiliser à la fois l'information fréquentiel et spatiale, d'autres méthodes ont été proposer, qui s'articulent généralement sur l'utilisation d'une fenêtre glissante qui se déplace dans l'image et de calculer l'énergie de la DFT de chaque fenêtre. Le choix de la taille de la fenêtre et du pas de déplacement dépend des caractéristiques spatiales des textures à analyser.

Parmi ces méthodes on cite :

- La transformée de Gabor.
- La transformée en ondelettes.
- La transformée de Huang-Hilbert.
- La transformée de Hermite.

I.3.1.2.2 Transformée de Gabor

Gabor définit en 1946 [9], pour appliquer les lois de la mécanique quantique pour mesurer le contenu de l'information temps-fréquence, il propose de décomposer un signal à partir d'ondes atomiques élémentaires. Pour cela la transformée de Fourier est légèrement modifiée par l'introduction d'une fenêtre glissante g , et devient :

$$\psi_{\lambda}(x, y) = g(x - x_0, y - y_0) \cdot e^{i(2\pi(u_0x + v_0y) + P)}$$

avec $\lambda = (x_0; y_0; u_0; v_0)$, $(u_0; v_0)$ définissant la fréquence spatiale de la sinusoïde, et P sa phase.

Lorsque g prend une forme gaussienne on dit qu'on effectue une analyse de Gabor. La fenêtre g représente alors une enveloppe gaussienne qui peut s'écrire :

$$g(x, y) = K e^{-\pi \left(\left(\frac{(x-x_0)\cos\theta + (y-y_0)\sin\theta}{\sigma_x} \right)^2 + \left(\frac{(y-y_0)\cos\theta - (x-x_0)\sin\theta}{\sigma_y} \right)^2 \right)}$$

Avec $(x_0; y_0)$ le sommet de la fonction ; σ_x et σ_y les écart-types de l'enveloppe gaussienne, K l'échelle de l'amplitude de l'enveloppe gaussienne, θ l'angle de rotation de l'enveloppe gaussienne. La fonction complexe de Gabor s'écrit alors :

$$\psi_{\lambda}(x, y) = K \cdot e^{-\pi \left(\left(\frac{(x-x_0)\cos\theta + (y-y_0)\sin\theta}{\sigma_x} \right)^2 + \left(\frac{(y-y_0)\cos\theta - (x-x_0)\sin\theta}{\sigma_y} \right)^2 \right)} \cdot e^{j(2\pi(xu_0 + yv_0 + P))}$$

En faisant varier les paramètres $\lambda = (x_0; y_0; u_0; v_0)$, plusieurs filtres peuvent être obtenus.

Enfin Les attributs caractérisant la texture de l'image sont finalement obtenus en calculant l'énergie, l'entropie ou la variance de chaque image filtrée.

Les filtres de Gabor sont très utilisés en analyse d'image et notamment dans la caractérisation et la segmentation des textures, et selon Fogel et Sagi [11] les filtres de Gabor ont de bonnes qualités à approximer les caractéristiques de cellules du cortex visuel des mammifères.

I.3.1.2.3 Transformée en ondelettes discrète

La transformée du Gabor se base sur des fenêtres de dimension fixe, ce qui n'est pas tellement parfait pour l'analyse des textures qui peut avoir des longueurs varier.

Pour pallier à cette handicap, la transformée en ondelettes consiste à l'utilisation de fenêtres de différentes taille pour l'analyse de l'image.

Tout comme les filtres de Gabor, les ondelettes permettent une bonne approximation du système visuel des mammifères. Mallat [10] affirme qu'elles permettent une analyse plus proche du système psycho-visuel par rapport à celle obtenue avec la famille de Gabor.

Mallat [10] a proposé en 1989 un algorithme pratique pour décomposer un signal et sa représentation en ondelettes, cet algorithme se base sur l'application successive de deux filtres, un filtre passe haut $h(x)$ et un filtre passe bas $g(x)$ sur l'image selon les lignes suivi d'un sous échantillonnage par un facteur de deux selon les lignes, puis en applique sur chacun du deux échantillons les deux filtre $h(x)$ et $g(x)$ selon les colonnes suivi par échantillonnage selon les colonnes. La figure I.4 présente la décomposition d'un signal 2D à partir de ces deux filtres. Quatre images résultats (A1, D1, D2, D3) sont obtenues ceci constitue un niveau de décomposition. A1 correspond aux basses fréquences horizontales et verticales, D1 aux hautes fréquences verticales et basses fréquences horizontales, D2 aux basses fréquences verticales et hautes fréquences horizontales, D3 aux hautes fréquences verticales et horizontales. L'image approximation A1 subit le même schéma de décomposition et ainsi de suite.

Les attributs de texture sont alors donnés par l'énergie des images détails de chaque niveau de décomposition.

L'avantage de cette transformée est qu'elle permet une caractérisation multi-échelles de la texture en considérant à la fois les informations globales et locales contenues dans l'image.

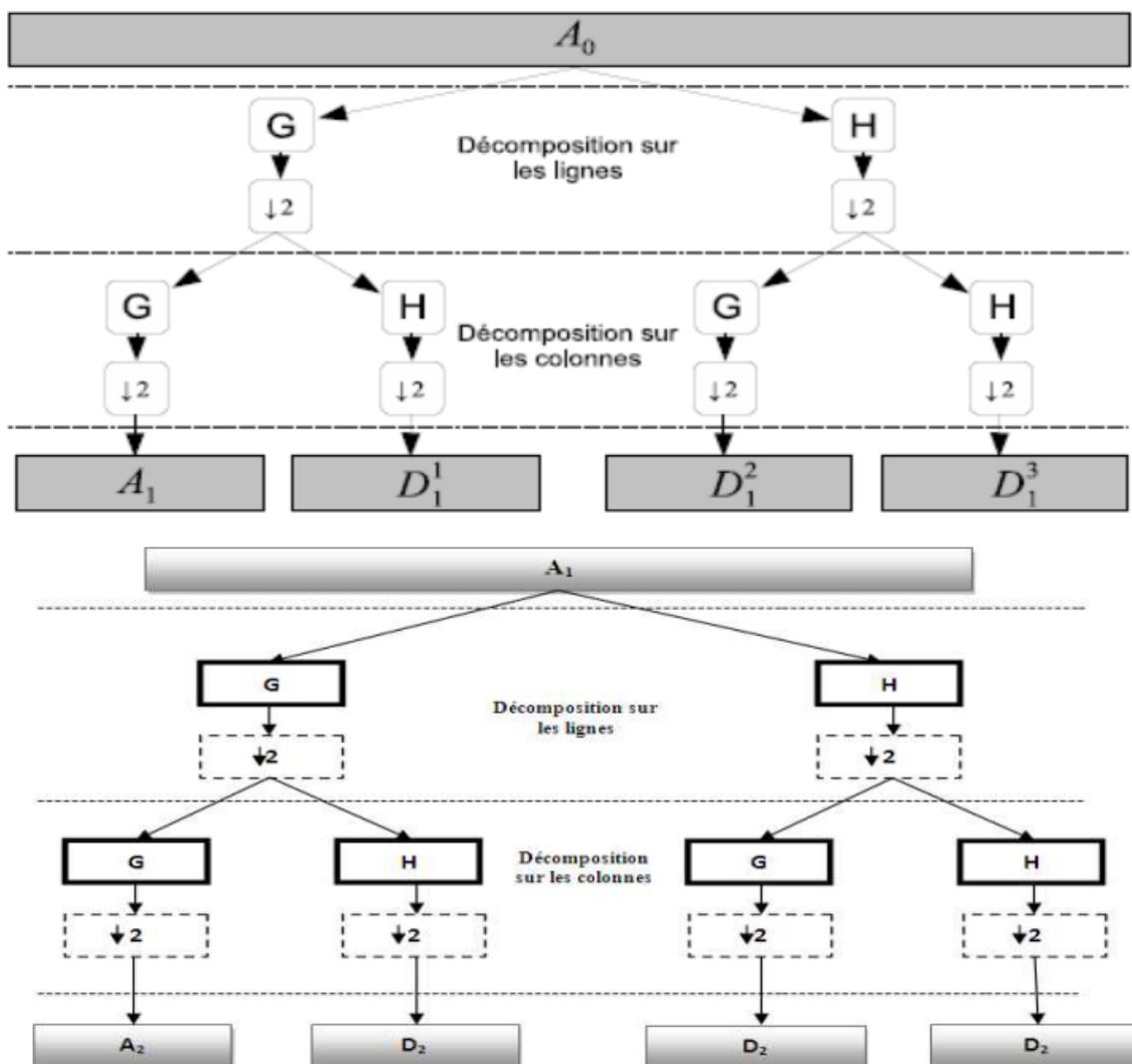


Figure I.4 : Décomposition en ondelette d'une image 2D

I.3.1.2.4 Frames ondelettes

Unser en 1995 [13] a proposé l'utilisation d'une méthode variante dite frame ondelette on éliminant l'étape de sous-échantillonnage de la décomposition en ondelette classique, et les résultats qu'il présente sont meilleurs qu'avec une décomposition en ondelette standard. Il fait également la comparaison avec des méthodes d'analyses de textures classiques comme les matrices de cooccurrences, les champs aléatoires de Markov Gaussien et des transformées linéaires locales, qui s'avèrent être moins performantes. Les frames d'ondelette permettent une description invariante aux translations dans une texture. Néanmoins cette méthode a une complexité de calcul non-négligeable ce qui peut être problématique dans certains cas d'applications.

I.3.1.2.5 Les paquets d'ondelettes

Une autre extension de la méthode de décomposition en ondelette standard et qui permis une analyse plus précise des fréquences dans une image est les paquets d'ondelettes. Dans cette extension le passage à un niveau inférieur se fait en décomposant les coefficients d'approximations ainsi que les coefficients de détails. Cette méthode permet d'avoir à la fois une analyse sur les basses fréquences mais aussi sur les hautes fréquences.

Laine et Fan en 1993 [14] ont utilisé cette approche. Les auteurs proposent une décomposition en paquets d'ondelette et calculent l'énergie en chaque sous-bande pour former un vecteur de caractéristiques afin de classifier des textures à l'aide d'un réseau de neurones. Laine et Fan [15] utilisent une extension de cette technique et la comparent à la méthode des frames d'ondelette afin de segmenter des textures de mosaïques. A partir d'un algorithme de classification non-supervisé (K-means), les résultats obtenus s'avèrent de meilleure qualité avec leur méthode en décomposition par paquet.

I.3.1.2.6 Le banc de filtres de Schmid

Schmid en 2001 [16] propose d'utiliser un ensemble d'opérateurs dit générique afin d'analyser les structures et les textures d'une image. L'ensemble de Schmid se compose de 13 filtres invariants à la rotation :

$$F(x, y, \sigma, \tau) = F_0(\sigma, \tau) + \cos\left(\frac{\sqrt{x^2 + y^2}\pi\tau}{\sigma}\right)e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Avec (σ, τ) prenant les valeurs suivantes : $(2,1)$, $(4,1)$, $(4,2)$, $(6,1)$, $(6,2)$, $(6,3)$, $(8,1)$, $(8,2)$, $(8,3)$, $(10,1)$, $(10,2)$, $(10,3)$ et $(10,4)$. τ représente le nombre de cycles de la fonction harmonique à l'intérieur de l'enveloppe Gaussienne du filtre.

I.3.1.2.7 Transformée de Huang-Hilbert

La transformée de Huang-Hilbert (THH) qui a été introduite pour la première fois en 1998 par Norden Huang et al. [17] est une autre méthode d'analyse temps fréquence.

Cette méthode est la combinaison de deux Transformées : la transformé de Huang (EMD) et la transformée de Hilbert (TH) pour l'estimation de l'amplitude instantanée et la fréquence instantanée.

L'EMD (Empirical Mode Decomposition) consiste à décomposer de façon adaptative le signal ou l'image en une somme de composantes oscillantes (IMFs) qui possèdent une seule fréquence (monomodale). La fréquence et l'amplitude instantanée (FI, AI) de chaque IMF sont ensuite calculées en utilisant la transformée de Hilbert. la figure 5 résume le principe de la transformée de Huang-Hilbert.

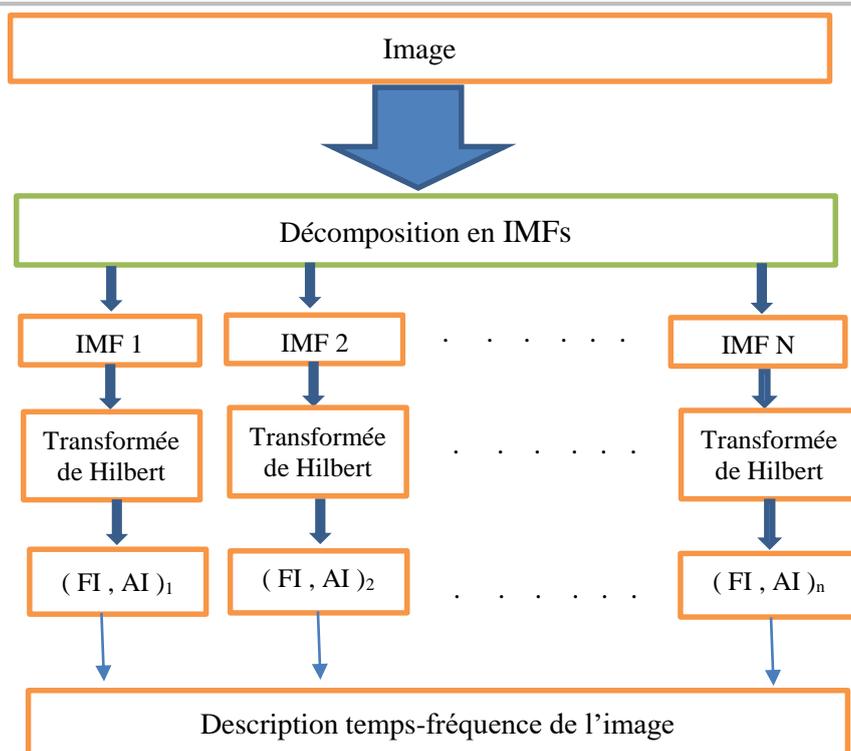


Figure I.5: Principe de la transformée de Huang – Hilbert.

I.3.2 Les méthodes basées sur des modèles

Le principe de ces méthodes est d'obtenir un modèle générateur de la texture. L'estimation des paramètres de ces modèles permet de caractériser les textures d'une image. On distingue en général trois types de modèles. Les modèles autorégressifs, les modèles fractals et les modèles Markoviens.

I.3.2.1 Les champs de Markov

Les champs de Markov sont des outils très utilisés en traitement d'image. Leur intérêt est principalement de pouvoir modéliser des interactions entre pixels de façon simple, à travers un modèle probabiliste. Considérons un champ aléatoire X constitué de variables aléatoires X_s en chaque site $s \in S$ (site s ici est le pixel). X^S est le champ X sans X_s , V_s désigne le voisinage de s . Les lettres majuscules désignent les variables et champs aléatoires, les lettres minuscules les valeurs qu'ils prennent, Un champ de Markov X est un champ aléatoire présentant la propriété suivante [18][19]:

$$P(X_s = x_s / X^S = x^S) = P(X_s = x_s / \{x_t\} \in V_s)$$

Les champs de Markov peuvent servir pour la modélisation des images, en particulier, des textures fines [20][21]. Dans ce cadre, il existe de nombreux modèles markoviens de texture, comme par exemple le modèle auto-binomial [22], le modèle de Derin-Elliott [23] ou le modèle markovien gaussien aussi appelé auto-normal[24].

I.3.2.2 Les fractales

Les fractales ont été utilisées avec beaucoup de succès pour synthétiser les textures naturelles. Mandelbrot [25][26] est le premier à relater l'existence de la géométrie fractale dans le monde. Un concept important concernant les fractals est l'autosimilarité. Un objet invariant par la transformation d'un vecteur \mathbf{r} est dit auto affine de vecteur \mathbf{r} . De plus, si les composantes du vecteur \mathbf{r} sont égales, alors nous pouvons parler d'autosimilarité.

La courbe de Koch [26] représente l'une des premières courbes fractales à avoir été décrite. Sa vecteur d'auto-affinement $r = (3,3)$. Elle a été inventée en 1906 par le mathématicien suédois Helge von Koch.

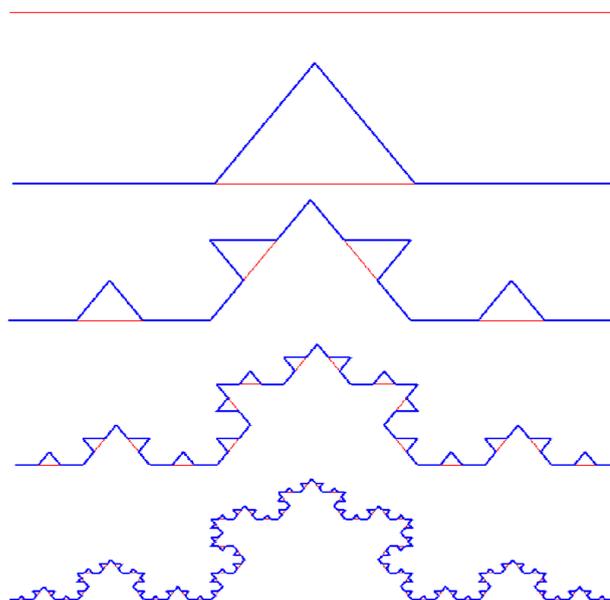


Figure I.6: Construction de la courbe de Koch.

Soille et Rivest ont publié en 1996 [27] un article dans lequel ils proposent un état de l'art sur les méthodes d'estimation de la dimension fractale. Celles-ci peuvent être divisées en deux familles comprenant les méthodes d'estimations géométriques et les méthodes stochastiques.

- Les méthodes stochastiques se base généralement sur l'analyse du spectre de puissance obtenu à partir de la transformée de Fourier [27], ou également en utilisant les variogrammes [32].
- Concernant les méthodes d'estimations géométriques, nous pouvons citer :
 - les méthodes d'analyse morphologique [28].
 - la méthode des triangles plans [29].
 - la méthode de l'élément structurant plat.
 - la méthode dite de « Box counting » [30][31].

La méthode du « Box counting » est l'une des plus utilisées grâce à sa simplicité de mise en œuvre ainsi que pour son faible coût en temps d'exécution. L'image est transformée en format binaire puis découpée en un quadrillage de carrés élémentaires de taille ϵ . Par la suite, on compte le nombre N_ϵ de carrés nécessaires à l'approximation de la texture. C'est à dire calculer la somme des boîtes contenant au moins un pixel noir. Le box dimension (D) d'une zone considérée X est alors donnée par :

$$D(X) = \lim_{\epsilon \rightarrow 0} \frac{\log[N_\epsilon(X)]}{\log \epsilon}$$

Pour déterminer cette limite, on considère plusieurs valeurs de ϵ . La pente de régression linéaire de la courbe ainsi obtenue correspond à la dimension fractale.

La dimension fractale n'est pas suffisante pour saisir toutes les propriétés texturales. Il a été montré qu'il peut y avoir des textures perceptuellement très différentes qui ont des dimensions fractales très similaires [33]. Par conséquent, une autre mesure, appelée lacunarité [26][30], a été suggérée afin de capturer la propriété texturale qui permettra de distinguer ces textures. Il est possible de l'estimer de la façon suivante :

$$\Lambda = \frac{E[M^2(\epsilon)] - E^2[M(\epsilon)]}{E^2[M(\epsilon)]}$$

Avec E l'espérance mathématique, M la masse fractale définie par :

$M(\epsilon) = \sum_{m=1}^{\epsilon^2} P_f(m, \epsilon) P_f(m, \epsilon)$ est la probabilité de trouver un point de l'objet fractal f dans un carré de taille ϵ .

I.3.3 Les méthodes basées sur les statistiques

Ces méthodes ne cherchent pas à identifier un motif de texture dans l'image, mais plutôt procède à l'évaluation statistique de la distribution des niveaux de gris dans l'image pour extraire les propriétés de la texture.

I.3.3.1 Statistique du premier ordre

A partir de l'histogramme de l'image on calcule des attributs représentant la probabilité d'avoir tel ou tel intensité à tel ou tel endroit dans l'image. Parmi les attributs qu'on peut calculé on trouve :

- les moments d'ordre k : $\mu_k = \sum_n p(n)n^k$
- la moyenne des niveaux de gris : μ_1
- les moments centrés d'ordre k : $\widetilde{\mu}_k = \sum_n (n - \mu_1)^k p(n)$
- la variance : σ^2 (moment centré d'ordre 2).
- le biais : $\frac{\widetilde{\mu}_3}{\sigma^3}$

- l'aplatissement : $\frac{\tilde{\mu}_4}{\sigma^4} - 3$
- l'énergie : $\sum_n |p(n)|^2$
- l'entropie : $-\sum_n p(n) \log_2 p(n)$
- le contraste : $\frac{\max(n) - \min(n)}{\max(n) + \min(n)}$
- la dynamique : $\max(n) - \min(n)$
- le coefficient de variation : $\frac{\mu}{\sigma}$

La combinaison de certaine ou de l'ensemble de ces attributs peut révéler des informations sur la texture présente dans l'image.

I.3.3.2 Caractéristiques d'autocorrélation

La fonction d'autocorrélation d'une image peut être utilisée pour évaluer le degré de régularité ainsi que la finesse / grossièreté de la texture présente dans l'image [2].

la fonction d'autocorrélation d'une image I de dimension $N \times M$ selon un vecteur de déplacement (x, y) est définie comme suit:

$$\rho(x, y) = \frac{\sum_{u=0}^N \sum_{v=0}^M I(u, v) I(u + x, v + y)}{\sum_{u=0}^N \sum_{v=0}^M I^2(u, v)}$$

Cette fonction est liée à la taille de la primitive de texture (c'est-à-dire à la finesse de la texture). Si la texture est grossière, la fonction d'autocorrélation décroît lentement ; sinon, sa valeur chute rapidement. Pour les textures régulières, la fonction d'autocorrélation affichera des pics et des creux.

I.3.3.3 Les Matrices de Cooccurrence

L'approche statistique du second ordre la plus connue pour l'analyse de la texture des images est la matrice de cooccurrence GLCM (Gray Level Cooccurrence Matrix). C'était le résultat de la première approche consistant à décrire, puis à classer, la texture de l'image. Il reste populaire aujourd'hui en raison de ses bonnes performances.

Parmi les propriétés de la texture on trouve la distribution et la cooccurrence spatiale de deux ou plusieurs niveaux de gris dans une région locale. Cette cooccurrence peut être résumée dans une matrice dont chaque élément est la probabilité (nombre d'apparition dans l'image) d'un couple de niveau de gris séparés par une relation spatiale R .

R est en général définie par un vecteur de déplacement $r(d, \theta)$ où d est la distance entre les deux pixels et θ la direction. Les matrices de cooccurrence étant symétriques, on les calculera en général dans les directions $0, 45, 90$ et 135° mais peuvent prendre également

d'autres valeurs compris entre 0 et 135°. [35]

Pour construire ces matrices M , $n \times n$ (avec n nombre de niveaux de gris de l'image), on va ainsi compter en $M(i, j)$, le nombre de fois où le niveau de gris j a succédé au niveau i pour un vecteur de déplacement $r(d, \theta)$.

Haralick [34] a proposé l'utilisation des matrices de cooccurrence pour extraire les caractéristiques de la texture de l'image. Pour une image codée en 256 niveaux de gris, sa matrice de cooccurrence est la matrice carrée de 256×256 tel que l'élément à la ligne i et la colonne j représente la probabilité (ou le nombre d'occurrence) qu'un pixel d'intensité i apparait dans l'image à une distance D et une direction θ d'un pixel d'intensité j .

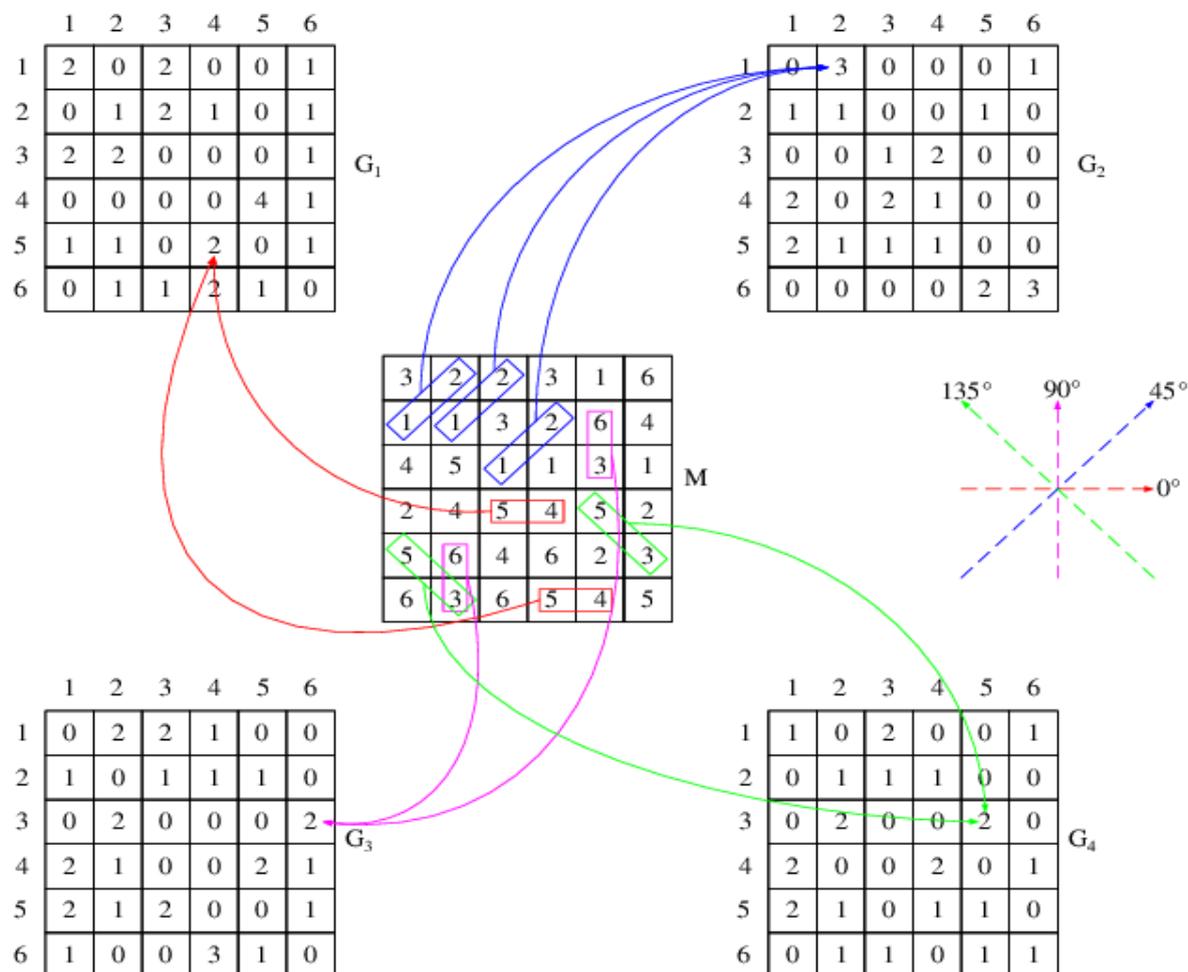


Figure I.7 : Formation de la matrice de cooccurrence GLCM.

La taille de la matrice de cooccurrence peut prendre de très grandes valeurs en dépendance avec le nombre de niveaux de gris de l'image, la matrice d'une image codée sur 8 bit (256 niveaux de gris) aura 256×256 éléments, ce qu'est pénalisant en terme de temps de calcul dès les résolutions plus élevées. Pour cette raison qu'on procède souvent à la quantification de l'image avant le calcul du GLCM.

Dans la plupart des applications, une seule matrice calculé qui est la somme des matrices obtenues pour $q = 0^\circ, 45^\circ, 90^\circ$ et 135° ; cette matrice dépend alors, plutôt du paramètre D . De plus, souvent, pour une direction donnée, on ne se préoccupe pas du sens de parcours, ce qui revient à rendre la matrice symétrique.

La matrice de cooccurrence contient un très grand nombre d'information, Ce qui les rend difficile à être exploiter directement, C'est pourquoi Haralick [34] a proposé d'en extraire 14 attributs qui décrivent les caractéristiques physique du texture. Ces 14 attributs sont :

Caractéristique de la texture	Description	Formule
1. L'énergie (ou moment angulaire second)	L'énergie mesure l'homogénéité de l'image. Plus cette valeur est faible, moins l'image est uniforme et l'image contient beaucoup de variation de couleur. (i , j) : coordonnées dans la matrice. P(i , j) : valeurs normalisées de la matrice. Valeur dans l'intervalle [0,1].	$f1 = \sum_i \sum_j P(i,j)^2$
2. Le contraste (ou inertie)	L'inertie mesure les variations locales des couleurs. Si ces variations sont importantes, alors le contraste sera élevé. Ce paramètre est fortement non corrélé à l'énergie. Valeur dans l'intervalle $[0, (Ng-1)^2]$.	$f2 = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (i-j)^2 P(i,j)$
3. Corrélacion	Où μ_x, μ_y, σ_x et σ_y sont respectivement les moyennes et les écarts type de p_x et p_y . Ce paramètre permet de déterminer si certaines colonnes de la matrice sont égales, c'est-à-dire s'il existe des dépendances linéaires dans l'image. En effet, plus les valeurs sont uniformément distribuées dans la matrice de cooccurrences et plus la corrélation est importante. Valeur dans l'intervalle [-1,1]. La corrélation n'est corrélée ni à l'énergie, ni à l'entropie.	$f3 = \frac{\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (ij)P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
4. La variance	La variance mesure la dispersion des valeurs autour de la moyenne. Plus ce paramètre est élevé et plus importants sont les écarts entre les valeurs et la moyenne. Elle représente l'hétérogénéité de la texture.	$f4 = \sum_i \sum_j (i - \mu)^2 P(i,j)$

<p>5. Moment différentiel inverse (Homogénéité)</p>	<p>Ce paramètre mesure l'homogénéité de l'image, plus l'image contient de régions homogène, plus il est élevé. Valeur dans l'intervalle [0,1].</p>	$f5 = \sum_i \sum_j \frac{P(i,j)}{1 + (i - j)^2}$
<p>6. Moyenne des sommes</p>		$f6 = \sum_{i=1}^{2Ng} i \cdot P_{x+y}(i)$
<p>7. Variance des sommes</p>		$f7 = \sum_{i=1}^{2Ng} (i - f8)^2 \cdot P_{x+y}(i)$
<p>8. Entropie des sommes</p>		$f8 = \sum_{i=1}^{2Ng} P_{x+y}(i) \cdot \log(P_{x+y}(i))$
<p>9. Entropie</p>	<p>Ce paramètre mesure le désordre dans l'image. sa valeur sera élevé en cas de texture complètement aléatoire (sans structure apparente). Lorsque les valeurs de la matrice de cooccurrences sont presque toutes égales, l'entropie est élevée.</p>	$f9 = - \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} P(i,j) \cdot \log(P(i,j))$
<p>10. Variance des différences</p>		$f10 = \sum_{i=1}^{Ng} (\mu - \mu_{x-y})^2 \cdot P_{x-y}(i)$
<p>11. Entropie des différences</p>	<p>Valeur dans l'intervalle [0,2.log(Ng-1)].</p>	$f11 = \sum_{i=1}^{Ng} P_{x-y}(i) \log P_{x-y}(i)$
<p>12. Information sur la corrélation 1</p>	<p>Soient :</p> <ul style="list-style-type: none"> • Px, Py les probabilités selon x et y, et HX, HY leurs entropie respective. 	$f12 = \frac{(HXY - HXY1)}{Max(HX, HY)}$
<p>13. Information sur la corrélation 2</p>	<ul style="list-style-type: none"> • $HXY = - \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} P(i,j) \log(P(i,j))$ • $HXY1 = - \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} P(i,j) \log(P_x(i)P_y(i))$ • $HXY2 = - \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} P_x(i)P_y(j) \log(P_x(i)P_y(i))$ 	$f13 = \sqrt{1 - e^{-2(HXY2-HXY)}}$
<p>14. Coefficient de corrélation maximal</p>	<p>Soient la matrice :</p> $Q(i;j) = \sum_k \frac{P(i,k)P(j,k)}{P_x(i)P_y(k)}$ <p>$V = 2^{eme}$ Plus grande valeur propre de Q</p>	$f14 = \sqrt{V}$

Tableau I.3 : Les 14 paramètres de Haralick.

Les attributs f_6 , f_7 , f_8 , f_{10} , f_{11} , f_{12} , f_{13} et f_{14} apportent des informations supplémentaires sur les degrés d'homogénéité et de complexité de l'image, ainsi que sur la corrélation.

La méthode d'extraction de ces paramètres basée sur le calcul des matrices de cooccurrences est une des méthodes les plus proches de la notion de texture. Elles mettent effectivement en avant les relations qui existent entre les pixels de l'image en faisant intervenir l'aspect local (les niveaux de gris) et l'aspect spatial (δ , θ). Cependant, ces matrices sont calculées de façon globale pour toute l'image, ce qui n'est pas sans poser de problème quand nous savons qu'il est possible d'avoir plusieurs textures différentes dans une même image.

Haralick [36] suggère de calculer des matrices de cooccurrence avec différentes directions et d'en faire la moyenne afin d'obtenir une seule matrice représentant les caractéristiques de la texture dans toutes les directions.

I.4 Conclusion

L'analyse de la texture est un domaine très important en traitement d'images numérique. Les méthodes d'analyse des textures sont riches et variées. Dans ce premier chapitre, nous avons présenté les méthodes d'analyse de texture d'image les plus connues ; en commençant par les définitions, ainsi que les types de textures en imagerie ; ensuite nous avons présenté les approches d'analyse de textures (fréquentiel, spatial, ou statistique), puis avons détaillé les différentes étapes pour le calcul des quatorze descripteurs de la méthode Haralick. Cette dernière que nous allons utiliser par la suite dans ce travail pour la classification des anomalies mammaires.

CHAPITRE II

MAMMOGRAPHIE

ET DEPISTAGE

DU CANCER DU SEIN

II.1. Introduction

Le cancer du sein représente plus d'un tiers des cancers féminins, le terme de cancer du sein ne désigne que les tumeurs malignes potentiellement agressives du sein tandis que le terme de « tumeur du sein » design à la fois les tumeurs malignes et les tumeurs bénignes.

Généralement, le cancer du sein apparaît sous forme de grosseur dans le tissu du sein, mais toutes les grosseurs ne sont pas cancéreuses. Ces grosseurs sont des tumeurs qui peut être maligne (cancéreuse) ou bénigne.

En Algérie comme à peu près partout dans le monde, le cancer du sein est le type de cancer le plus courant chez les femmes. Il affecte habituellement les femmes de plus de 50 ans [45]. Cependant, les femmes de tous âges peuvent avoir un cancer du sein. Dans la plupart des pays du monde, Toutes les femmes qui sont entre 50 et 70 ans devraient être régulièrement examinées dans le cadre du programme de dépistage du cancer du sein. Les femmes plus jeunes ayant un risque élevé de cancer du sein peuvent aussi bénéficier du dépistage. Il est très important que les femmes examinent régulièrement leurs seins pour détecter toute anomalie, et qu'elles fassent examiner tout changement par leur médecin.

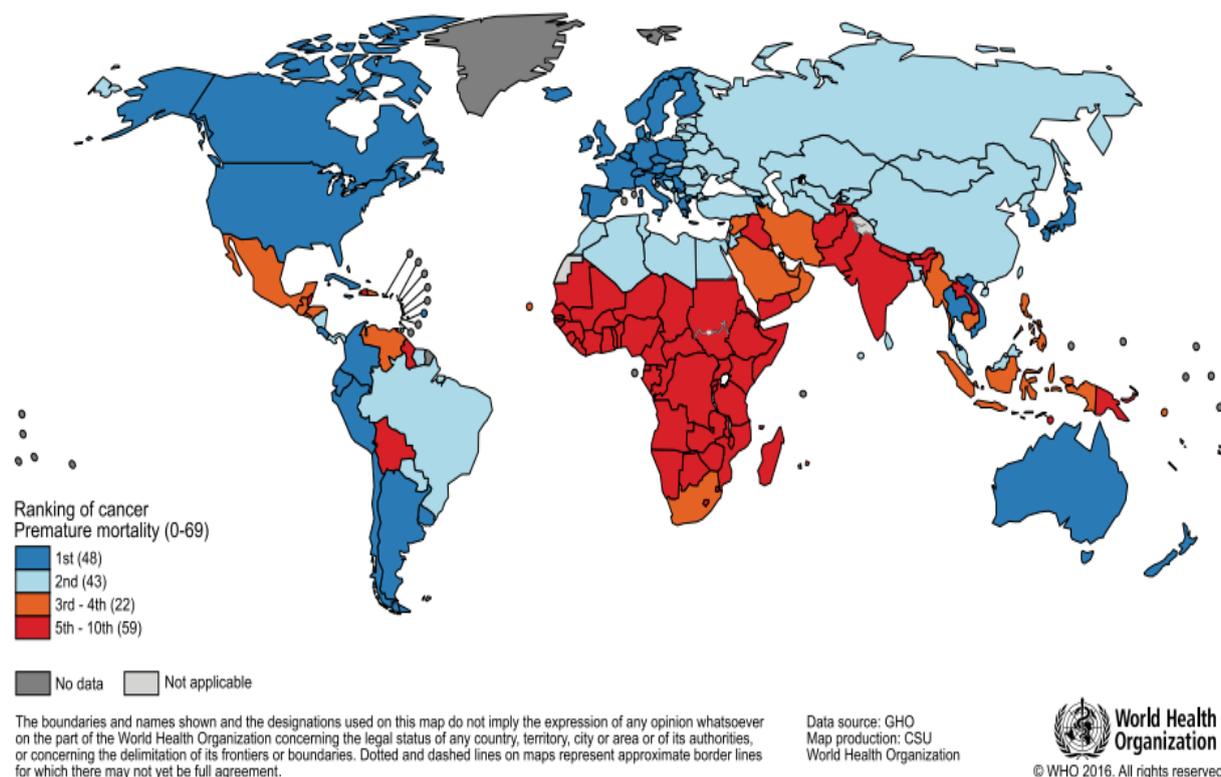


Figure II.1: Carte mondiale présentant le classement national du cancer comme cause de décès pour les âges inférieurs à 70 ans en 2015 [44]

II.2 Anatomie du sein

Le sein est un organe globuleux et pair contenant la glande mammaire, Il contient de la graisse, des tissus conjonctifs et de nombreuses petites glandes qui produisent du lait. Chez les femmes allaitantes, le lait atteint le mamelon en passant par de petits tubes appelés des canaux. Les 2 seins occupent la partie antéro-supérieur du thorax, plus ou moins symétrique, devant les muscles pectoraux, face à l'espace entre les arcades antérieures de la troisième et de la septième côte.

Il est constitué de tissus adipeux sous-cutanés, des tissus conjonctifs gras, des canaux galactophores qui achemine le lait au mamelon, et de nombreuses petites glandes (Lobules) qui produisent du lait.

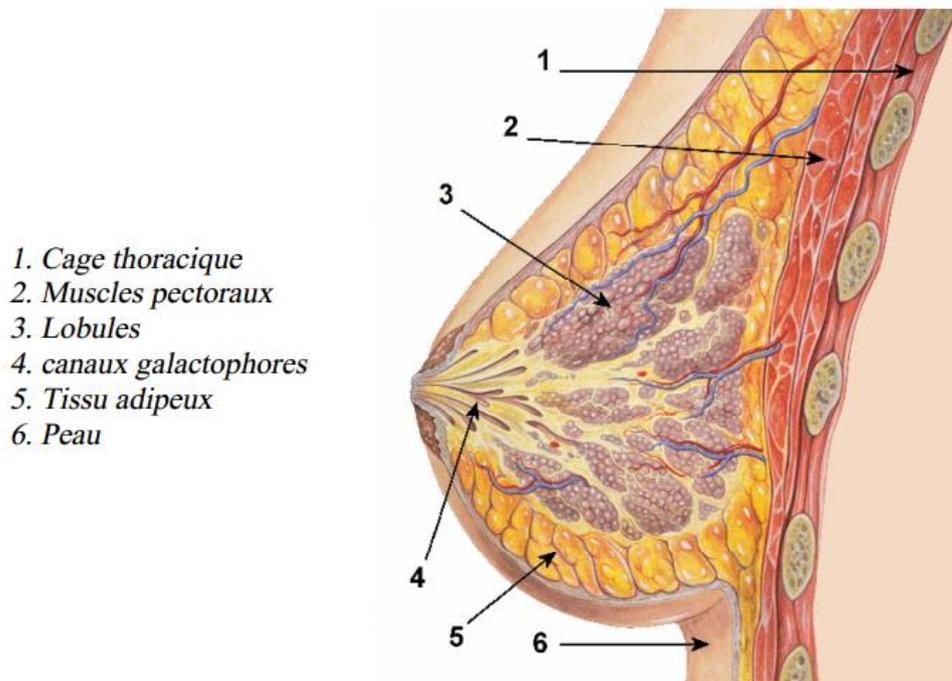


Figure II.2 Schéma anatomique du sein

II.3 Technique de dépistage du cancer de sein

Le cancer du sein est connu depuis l'Antiquité [46], car nous trouvons des descriptions de la maladie dans les écrits de médecins grecs et égyptiens antiques il y a plus de 3500 ans. Jusqu'à la moitié du 19ème siècle, la chirurgie était la seule l'option disponible pour le cancer du sein.

En 1976, les résultats publiés de Fisher [47] utilisant une chirurgie de sein plus simple suivi de la radiothérapie ou de la chimiothérapie. Il a noté que c'étaient juste comme efficaces que la mastectomie radicale.

Puis dans les années 80 et les débuts des années 90, les traitements du cancer s'est vite développé, et des technique de dépistage précoce on vit le jour telle que la mammographie.

Plusieurs pays en mis des programmes de dépistage et de suivi des femmes qui présente une forte probabilité de risque de développer le cancer du sein. Le programme consiste a pratiqué une mammographie toute les 2 ou 3 ans pour les femmes qui ont dépassé l'âge de 45-50 ans ou ceux qui ont un antécédent familiale de cancer.

Les trois méthodes principales utilisées soit pour le dépistage, ou le diagnostic, sont la mammographie, l'échographie, et l'imagerie par résonance magnétique (IRM).

II.3.1 La Mammographie

La mammographie est un examen réalisé avec un appareil numérique à rayons X (le mammographe). Le sein est un organe dont les différents constituants ont des coefficients d'atténuation aux rayons X presque identiques. Les différents composants mammaires ont un contraste peu élevé. La détection radiologique des lésions du sein exige une imagerie de haute résolution puisque l'on recherche la présence de micro calcifications. De telles images ne sont obtenues que par un appareillage ayant des caractéristiques spéciales. Le faisceau de rayons X est produit de manière à obtenir une imagerie de haute résolution. Le mammographe possède un dispositif de compression du sein. Cette compression du sein participe tout autant au processus permettant d'obtenir des images de haute résolution qu'à une diminution des doses de rayonnements utilisés. Aujourd'hui cette dose de rayonnement est si faible que la discussion sur la dangerosité de la mammographie n'est plus d'actualité. Hormis ces conditions de haute résolution et de contraste, le principe de formation de l'image est le même que pour une radiographie.

Cet examen permet de visualiser la structure du sein. Chaque sein est habituellement radiographié dans deux positions différentes de façon à obtenir une analyse de l'ensemble du volume de la glande mammaire. Dans le protocole standard du dépistage, une incidence cranio-caudale (CC) est une incidence médio latérale oblique (MLO) sont réalisées sur chaque sein. Le but de l'examen mammographie de dépistage est le diagnostic du cancer du sein à un stade précoce, permettant ainsi un traitement efficace et peu agressif. Les programmes organisés de dépistage du cancer du sein ont ainsi permis de faire baisser de manière significative la mortalité par cancer du sein dans les populations dépistées (25% à 48% [37][38]).

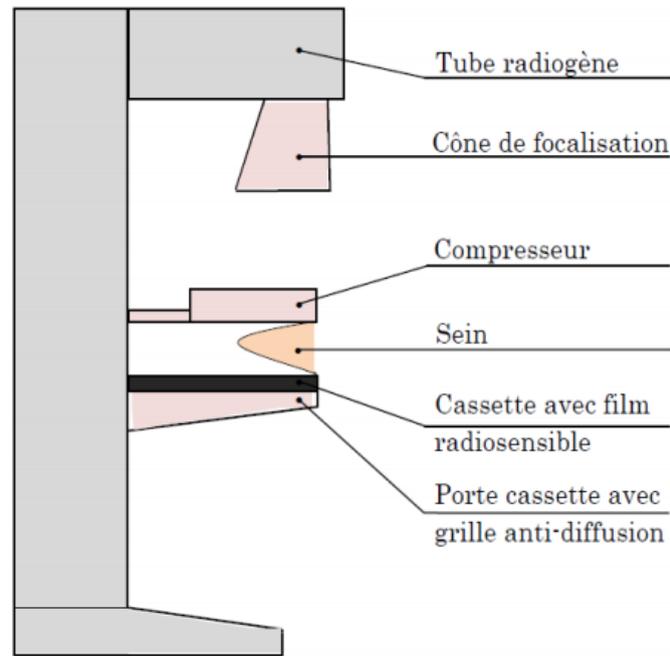
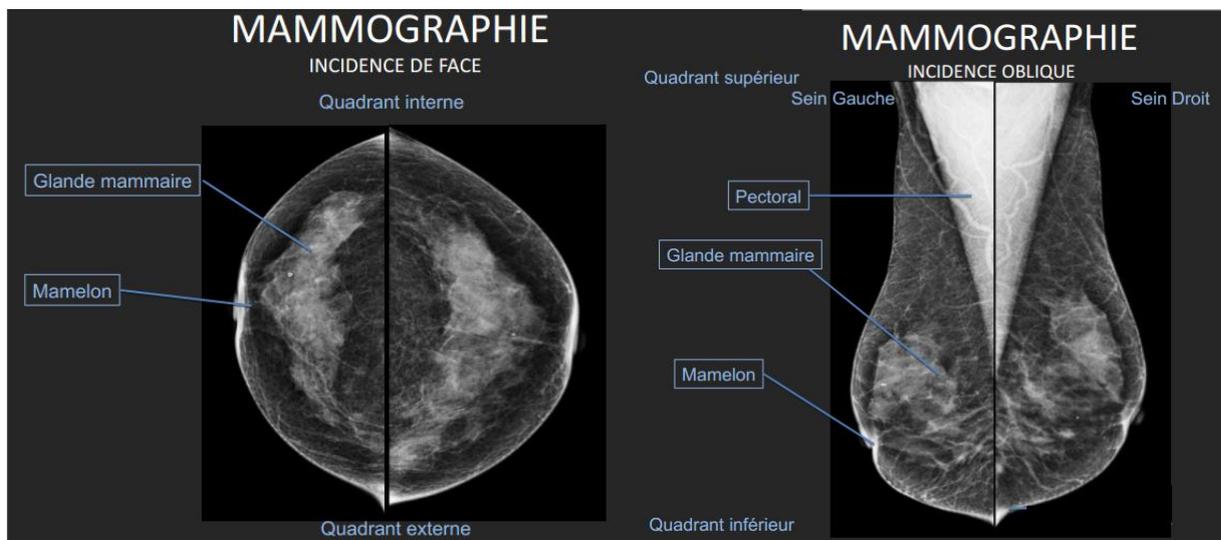


Figure II.3 Les composants d'un mammographe.



- A -

- B -

Figure II.4: A : Résultat de l'incidence de face, B : Résultat de l'incidence oblique.

II.3.1.1 Déroulement de l'examen de mammographie

La femme venant réaliser sa mammographie de dépistage est accueillie par la technicienne en radiologie médicale qui va procéder à la réalisation de la mammographie. Le sein est positionné sur le plateau du mammographe et une compression progressive est appliquée. Le tube à rayons X est activé durant quelques secondes et l'image du sein est obtenue. Cette manœuvre est répétée pour chaque sein dans les incidences CC et MLO [48].

Cette manœuvre, qui peut être quelquefois légèrement douloureuse, est indispensable pour obtenir un cliché de qualité permettant un diagnostic performant. Non seulement la compression améliore la définition des structures, mais elle permet de réduire la dose d'irradiation [48].

II.3.1.2 Limite de la mammographie

La mammographie ne permet pas seul de confirmer la présence ou l'absence du cancer, elle permet de voir s'il existe des anomalies dans le sein, pour déterminer avec certitude le résultat d'un examen mammographique, d'autres procédures et analyses seront pris en cas de détection d'anomalies [75].

Pour les adolescents ou les jeunes femmes qui n'ont pas encore des enfants, leurs seins contiennent plus de glande et peut de graisse. Or, la glande est mal étudiée par les rayons X et les mammographies sont donc peu adaptées et/ou insuffisantes pour cette catégorie de patientes [75].

II.3.2 L'échographie

L'échographie est une technique d'imagerie employant des ultrasons, une sonde placée sur la peau en glissant au moyen d'un gel, émet des ondes sonores de hautes fréquences, inaudibles, qui traversent le sein et retournent des échos, ces derniers interprétés par un ordinateur sont convertis en images, ce qui permet la visualisation de l'intérieur du sein.

Indolore et non invasive, elle est très largement utilisée dans le diagnostic de nombreuses pathologies, ainsi que dans le guidage visuel lors d'autres examens, comme la biopsie.

Mais l'inconvénient principal de cette méthode, est son incurabilité de détecter les anomalies de petite taille tel que les microcalcifications. Cela le rend très peu fiable et n'est utilisé que comme un examen complémentaire à une mammographie [76].

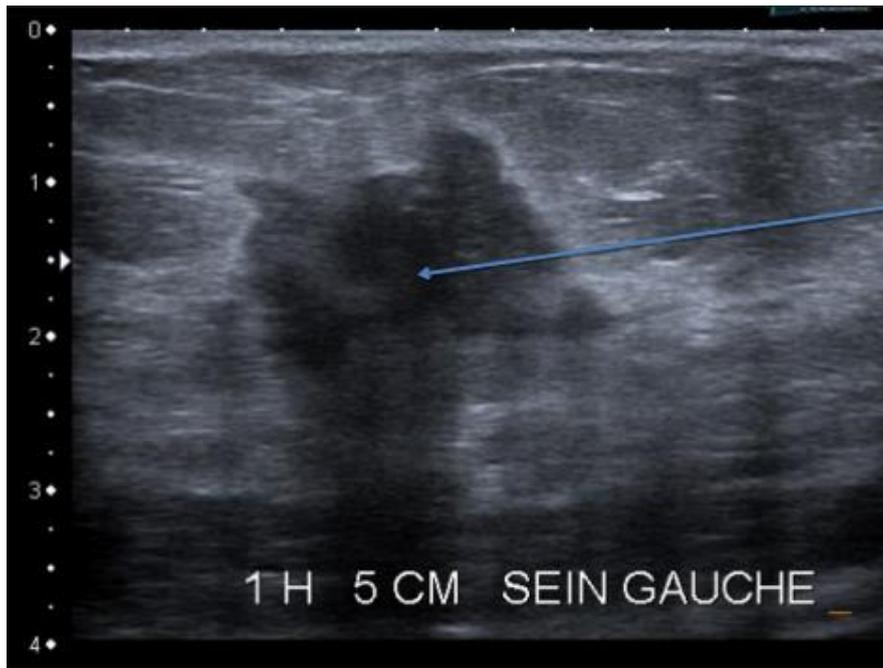


Figure II.5: *Echographie du sein pour diagnostic du cancer.*

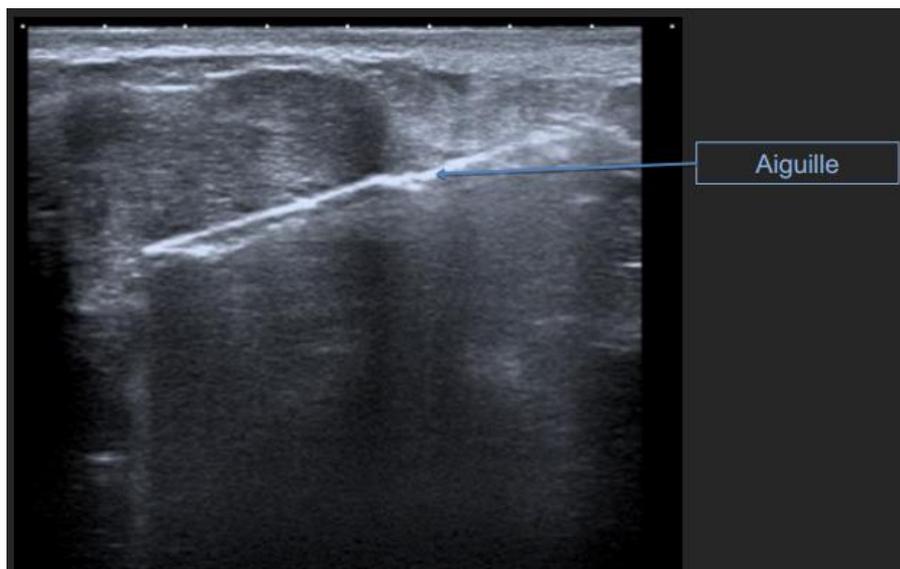


Figure II.6 : *Echographie du sein pour guidage lors d'une opération de biopsie.*

II.3.3 Imagerie par Résonance Magnétique (IRM)

L'imagerie par résonance magnétique (IRM) est un examen qui permet d'obtenir des vues en deux ou trois dimensions de l'intérieur du sein. L'IRM donne des informations sur des lésions qui ne sont pas visibles sur les radiographies standards, l'échographie ou le scanner.

Elle utilise un champ magnétique (aimant) et des ondes radio. Aucune radiation ionisante n'est émise. Son principe consiste à réaliser des images du corps humain grâce aux nombreux atomes d'hydrogène qu'il contient. Placés dans un puissant champ magnétique, tous les atomes d'hydrogène s'orientent dans la même direction : ils sont alors excités par des ondes radio durant une très courte période (ils sont mis en résonance). A l'arrêt de cette stimulation, les atomes restituent l'énergie accumulée en produisant un signal qui est enregistré et traité sous forme d'image par un système informatique [76].

L'IRM étudie avec une grande précision de nombreux organes tels que le cerveau, la colonne vertébrale, les articulations et les tissus mous. Elle est d'une grande utilité lorsqu'une analyse très fine est nécessaire et que certaines lésions ne sont pas visibles sur les radiographies standards, l'échographie ou le scanner. Elle permet de faire des images en coupes dans différents plans et de reconstruire en trois dimensions la structure analysée.

L'IRM du sein ne remplace pas la mammographie ni l'échographie mammaire [76]. Elle ne constitue pas un examen systématique du diagnostic du cancer du sein. Il s'agit d'un outil supplémentaire qu'on emploie parfois :

1. Pour faire la différence entre une anomalie bénigne et une anomalie cancéreuse dans le cadre du dépistage des femmes à haut risque de cancer du sein ;
2. Lorsque l'imagerie standard (mammographie ou échographie) ne permet pas de conclure avec certitude à l'absence de malignité ;
3. Pour guider un prélèvement par biopsie ;
4. Dans le cadre du bilan d'extension ;
5. Pour évaluer la réponse aux thérapeutiques néo-adjuvantes (chimiothérapie réalisée avant l'opération) ;
6. A la recherche de récurrence locale après la chirurgie conservatrice ;
7. Pour vérifier l'état d'implants mammaires, en particulier si on croit qu'il y a eu rupture.

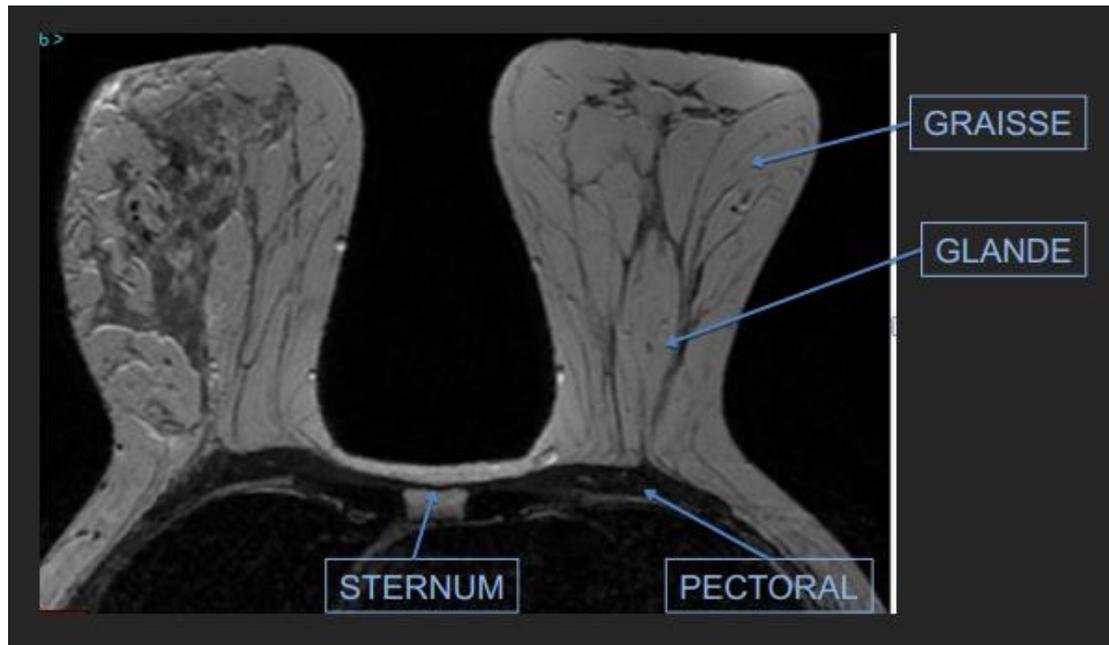


Figure II.7 : IRM mammaires coupe axiale.

II.4 Types d'anomalies détectables dans un dépistage de cancer de sein

Une anomalie correspond à une plage de surdensité anormale détecté sur l'une des deux cliché du sein, il existe aussi beaucoup de surdensités normale, La distinction entre les deux est due à l'expérience du spécialiste on analysant plusieurs caractéristiques comme la taille, la densité, les contours, la forme, les texture.

On peut classer les anomalies selon trois catégories :

- 1) Les densités asymétriques.
- 2) Les masses.
- 3) Les calcifications.

II.4.1 Les densités asymétriques

Une densité asymétrique est détecter en plaçant la radiographie des seins gauche et droit l'un à côté de l'autre comme un miroir, et en cherche les différences de densité entre les deux seins. Dans le cas des seins normaux, la répartition des densités est symétrique, dans le cas contraire on observe une asymétrie de densités d'un sein à l'autre ce qui peut être signe d'une tumeur déformant le sein [39].

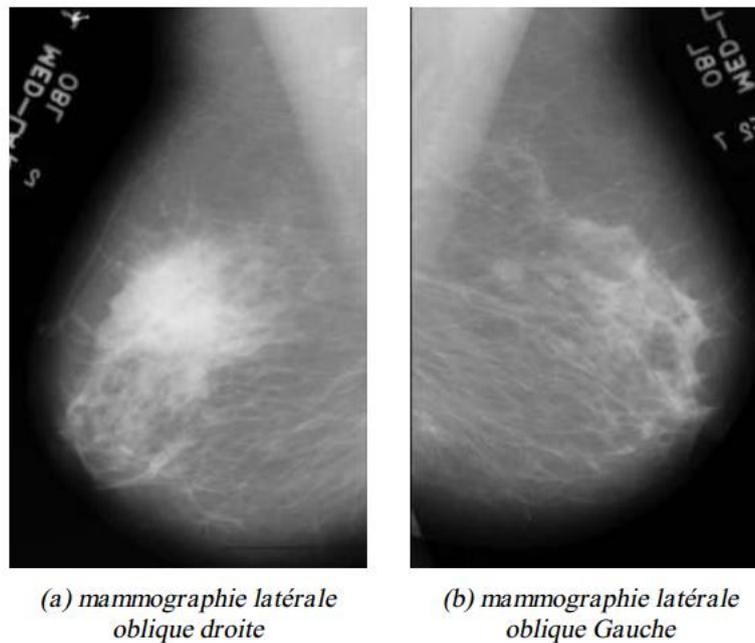


Figure II.8: une asymétrie de densité observée seulement sur la mammographie gauche.

II.4.2 Les masse

Une masse est une opacité tridimensionnel ayant un volume relativement grand, perceptible depuis les différentes incidences, l'analyse de ces caractéristiques tel que le contour et la forme révèle son malignité, On distingue généralement 5 formes différentes de masses : ronde, ovale, lobulaire, ou irrégulière [40].

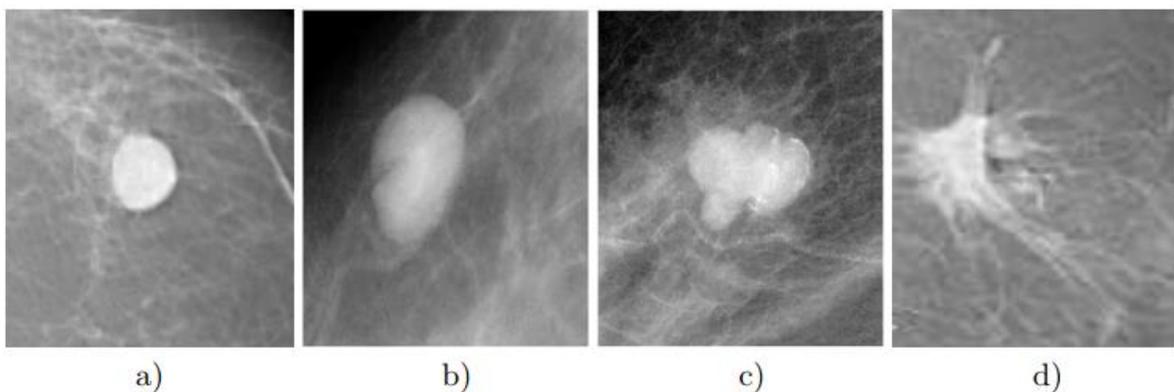


Figure II.9 : Les différentes formes possibles d'une masse :

a) Ronde, b) Ovale, c) Lobulée et d) Irrégulière

Concernant les contours, leurs aspect est un indice signifiant dans la malignité ou non de la masse, On distingue généralement 5 types de contours : bien définis, effacés, micro-lobulés, mal définis et avec des spicules.

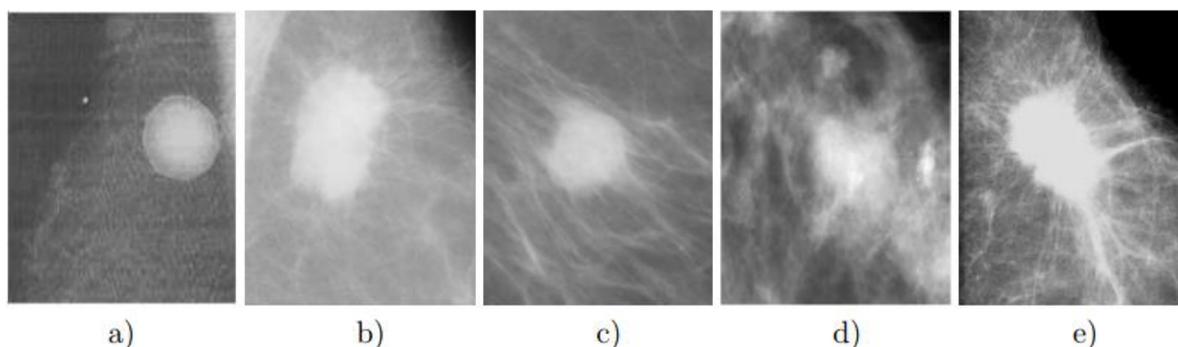


Figure II.10 : Les différents contours possibles d'une masse : a) Bien définis, b) Micro-lobulé, c) Effacés, d) Mal définis et e) Spiculé

Les distorsions architecturales (figure II.11-A) et les masses avec des contours spiculés (figure II.11-B) représentent une forte probabilité de présence de cancer.

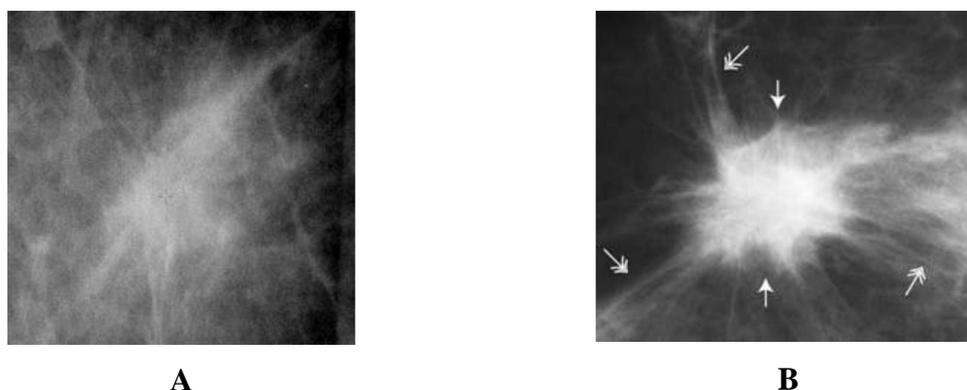


Figure II.11 : A- Images de distorsion architecturale. B-Images de masses avec des limites en spicules.

II.4.3 Les calcifications

Les calcifications mammaires sont des dépôts de calcium qui se forment dans le tissu du sein. Elles sont assez fréquentes et la plupart ne sont pas liées au cancer. Elles ne sont pas liées non plus à la quantité de calcium pris dans l'alimentation ou les suppléments prenaient.

On détecte souvent les calcifications mammaires lors d'une mammographie de dépistage. Elles apparaissent sous la forme de points blancs sur les clichés mammaires. Le médecin observe la taille, la forme et la disposition des calcifications. Il pourrait décider d'approfondir l'évaluation de certaines caractéristiques des calcifications, comme une forme irrégulière ou le nombre de calcifications dans une région du sein [41].

Trois propriétés essentielles sont étudiées par le radiologue pour juger leur probabilité d'être bénigne ou maligne, le nombre, la taille et la forme.

Selon le nombre, au-dessus de 4 à 6 micro-calcifications le taux de soupçon de cancer augmente considérablement.

Selon la taille des grains de la calcification on distingue deux types : les macro-calcifications et les micro-calcifications.

II.4.3.1 Micro-calcifications

Les micro-calcifications sont de minuscules dépôts de calcium dans le sein. Une seule micro-calcification peut se développer ou bien plusieurs peuvent former une grappe dans une région du sein.

Quand les cellules du sein se développent et se divisent, elles fabriquent plus de calcium. C'est pourquoi les micro-calcifications apparaissent souvent dans une région du sein où les cellules se développent et se divisent. Lorsque les cellules sont plus actives dans une certaine région du sein, cela peut signifier qu'il y a des cellules cancéreuses dans cette région [41].

II.4.3.2 Macro-calcifications

Les macro-calcifications sont de gros dépôts grossiers de calcium dans le sein. Elles sont assez fréquentes chez les femmes âgées de plus de 50 ans. On les associe souvent à des changements non cancéreux (bénins) qui se produisent dans le sein et qui sont provoqués par les éléments suivants :

1. vieillissement des artères
2. lésions aux tissus du sein causés par un accident de voiture par exemple
3. traitements antérieurs d'un cancer du sein, dont la chirurgie et la radiothérapie
4. infection des tissus du sein (mastite)
5. masses non cancéreuses au sein telles qu'un fibroadénome ou des kystes
6. dépôts de calcium dans la peau ou les vaisseaux sanguins

La forme des micro-calcifications révèle beaucoup sur leurs malignités, Par exemple, les calcifications arrondies ou ovales et d'une taille uniforme, sont probablement bénignes. Au contraire, lorsqu'elles sont irrégulières dans leur taille, et si elles sont représentées par leur hétérogénéité comme « des débris de verre », sont probablement malignes [41]. On distingue 5 types de calcifications.

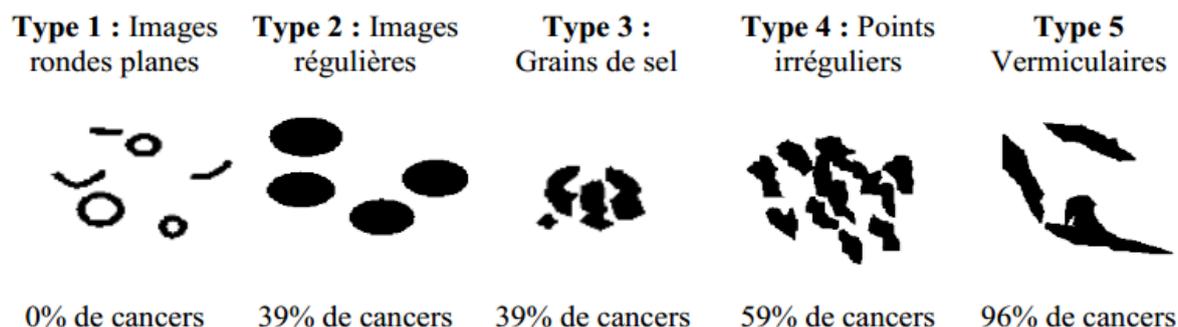


Figure II.12 : Type de calcification.

II.5 Classification des anomalies mammaires

Afin de parler le même langage par toute les radiologues et médecins sensé lire et interpréter les rapports de mammographie, une nécessité d'adopté une classification des anomalies compris par tout le monde est apparait tôt après l'application de cette technique de mammographie. Dans ce cadre Il existe plusieurs classifications comme Le Gal [42], de Lanyi [44] et de BIRADS [43].

II.5.1 La classification de LeGal

Le Gal en 1976 et 1984 [42] présentait une classification radiologique basée sur l'aspect radiologique des micro-calcifications (disposition, régularité des contours, des formes, des tailles et des densités) avec une évaluation croissante de la prédictivité de malignité.

Type Le Gal	Aspects mammographiques	Interprétation
Type 1	Micro-calcifications annulaires rondes à centre clair radio-transparentes.	Elles correspondent dans tous les cas à une pathologie bénigne souvent des dépôts calciques stratifiés circulaires dans l'épaisseur de la paroi de galactophores dilatés.
Type 2	Micro-calcifications punctiformes, rondes, pleines, radio-opaques, aux contours réguliers et arrondis,	3 Dans 20% des cas, il s'agit de lésions malignes 4 dans 20% des cas, il s'agit de lésions frontières 5 dans 60% des cas, il s'agit de lésions bénignes.
Type 3	Micro-calcifications poussiéreuses, trop fines pour préciser leur forme.	Des lésions bénignes (50%) ; des lésions malignes (50%)
Type 4	Micro-calcifications punctiformes irrégulières aux contours anguleux,	Dans 70% des cas, il s'agit de lésions malignes
Type 5	Micro-calcifications vermiculaires, elles ont la forme d'un bâtonnet souvent irrégulier.	100 % de lésions malignes ; dans la majorité des cas, il s'agit d'un carcinome.

Tableau II.1 : Classification de Le Gal

Cette classification admise et actuellement encore très utilisée, elle a l'avantage d'être simple, toutefois elle présente quelques points faibles en particulier :

1. Ne tient pas compte de la forme et du nombre des foyers de micro-calcifications.
2. L'étude se fait uniquement sur les micro-calcifications, et ignore les autres types de lésion telle que les masses et distorsion architecturales.
3. L'étude des signes associés.

II.5.2 La classification de Lanyi

Lanyi [44], à partir des travaux sur le mécanisme de formation des micro-calcifications, a proposé une classification basée sur la théorie du moule : la forme des micro-calcifications dépend de l'endroit où elles ont été formées. Il distinguait ainsi les micro-calcifications rondes, bénignes à point de départ lobulaire et les micro-calcifications allongées, associées à des lésions malignes d'origine galactophorique. Malheureusement, certaines micro-calcifications avec forme arrondie sont associées à des cancers : carcinome lobulaire in situ, carcinome canalaire cribriforme, carcinome tubuleux. Certaines pathologies bénignes se manifestent avec des micro-calcifications de forme allongée : certaines calcifications débutantes d'un adéno-fibrome, lithiases galactophoriques, certaines mastopathies complexes. Lanyi a également insisté sur l'intérêt de l'étude de la forme, du nombre des foyers de micro-calcifications ainsi qu'au polymorphisme des micro-calcifications au sein d'un même foyer.

II.5.3 La classification de BI-RADS

Pour pallier aux lacunes de la classification de Le Gal, l'American College of Radiology (ACR) a souligné l'importance d'un protocole mammographique standardisé et complet qui tient en compte les différents facteurs de malignité. En novembre 1998, l'ACR a établi le système Américain BI-RADS (Breast Imaging Reporting and Data System) qui a été rédigé par un groupe d'experts réunis par l'ANAES (Agence Nationale d'Accréditation et d'évaluation en Santé) [45]. Ce système classe les anomalies mammographiques en six catégories en fonction du degré de suspicion de malignité,

La classification de l'ACR résume les formes des différentes masses et des différentes micro-calcifications, leur texture, les différents aspects de la distorsion architecturale ainsi que leur degré de malignité. En plus de l'identification du type lésion, le système BI-RADS donne aussi la conduite à suivre pour chaque classe, le tableau II.2 récapitule les classes du système, et le tableau II.3 donne la conduite à suivre pour chaque classe.

BI-RADS0		Cliché insuffisant pour vérifier la présence ou non de lésions
BI-RADS1	Masses	Absence de masses suspectes.
	Micro-calcifications	Absence de micro-calcifications suspectes.
BI-RADS2	Masses	<ul style="list-style-type: none"> - Opacité ronde avec macro-calcifications. - Ganglion intra mammaire. - Images de densité graisseuse ou mixte. - Opacités rondes correspondant à un kyste en échographie.
	Micro-calcifications	<ul style="list-style-type: none"> - Macro-calcifications cutanées. - Calcifications punctiformes régulières diffuses. - Calcifications sur matériel de suture. - Macro-calcifications sans opacité. - Micro-calcifications annulaires, semi-lunaires ou sédimentées.
BI-RADS3	Masses	- Masse circonscrite ronde ou ovale sans micro-lobulation non liquidienne en échographie.
	Micro-calcifications	<ul style="list-style-type: none"> - Micro-calcifications rondes ou punctiformes régulières ou pulvérulentes, peu nombreuses, en petit amas rond isolé. - Petit amas rond ou ovale de calcifications amorphes et peu nombreuses.
BI-RADS4	Masses	<ul style="list-style-type: none"> - Masse spiculée sans centre dense. - Masse ronde ou ovale à contour lobulé ou masqué.
	Micro-calcifications	<ul style="list-style-type: none"> - Micro-calcifications pulvérulentes groupées et nombreuses. - Micro-calcifications irrégulières, polymorphes ou granulaires, peu nombreuses. - Micro-calcifications punctiformes régulières nombreuses et/ou groupées en amas à contour ni rond, ni ovale.
BI-RADS5	Masses	<ul style="list-style-type: none"> - Masse spiculée à centre dense. - Masse à contour flou et irrégulier.
	Micro-calcifications	<ul style="list-style-type: none"> - Micro-calcifications associées à une anomalie architecturale ou à une opacité. - Groupement de Micro-calcifications quelle que soit leur morphologie, dont la topographie est galactophorique. - Micro-calcifications vermiculaires, arborescentes ou irrégulières polymorphes ou granulaires nombreuses et groupées. - Micro-calcifications groupées ayant augmenté en nombre ou Micro-calcifications dont la morphologie et la distribution sont devenues plus suspectes.
BI-RADS 6		Correspond à une mammographie ayant une malignité prouvée. Une action thérapeutique doit être entreprise (chimiothérapie, radiothérapie, hormonothérapie, intervention chirurgicale).

Tableau II.2 : Classification BIRADS de l'ACR.

BI-RADS0	Nécessité d'investigations complémentaires : telles que cliché avec compression centrée, agrandissement, incidence particulière, échographie.
BI-RADS1	Négative : mammographie normale, aucune masse ou calcification suspecte n'est présente.
BI-RADS2	Lésions bénignes : ne nécessitant ni surveillance ni examen complémentaire
BI-RADS3	Lésions probablement bénignes : une surveillance clinique et radiologique à court terme est conseillée.
BI-RADS4	Anomalies suspectes : une biopsie devrait être envisagée. Ces lésions n'ont pas un aspect typique de cancer, mais peuvent néanmoins correspondre à une lésion maligne.
BI-RADS5	Lésions fort suspectes de malignité : l'anomalie est évocatrice d'un cancer.
BI-RADS6	Malignité approuvé après une biopsie.

Tableau II.3 : Conduite à tenir pour chaque classe de l'ACR

II.6 Description de différentes bases de données mammographiques

Lors de l'analyse comparative d'un algorithme ou de caractéristiques, il est recommandé d'utiliser une base de données de test standard pour que les résultats des expériences puissent être directement comparés. La plupart des bases de données mammographiques ne sont pas accessibles au public. Les bases de données les plus facilement accessibles et donc les bases de données les plus couramment utilisées sont la base de données de la « **Mammographic Image Analysis Society** » (MIAS) et la « **Digital Database for Screening Mammography** » (DDSM). En outre, il existe actuellement peu de projets développant de nouvelles bases de données d'images mammographiques comme il y avait plusieurs projets anciens.

II.6.1 La base de données MIAS

La « **Mammographic Image Analysis Society (MIAS)** » est une organisation britannique de groupes de recherche intéressés par les études des mammographies, elle a généré une base de données de mammographies numériques. Les films tirés du programme national du royaume uni de dépistage de cancer des seins, ils ont été numérisés à une résolution de 50 microns, chaque pixel est codé sur 8 bits. La base de données contient 322 images numérisées. Il comprend également les marques du radiologue sur toute anomalie existante. La base de données a été réduite à une résolution de 200 microns et redimensionnés de sorte que toutes les images soient d'une taille de 1024x1024.

II.6.2 La base de données AMDI

AMDI (**Indexed Atlas of Digital Mammograms**) fournit un outil qui permet à l'utilisateur de télécharger des cas à partir de la base de données mammographiques, afin de mettre les informations à la disposition des communautés médicales et de recherche autorisées intéressées par le diagnostic du cancer du sein. La base de données mammographiques a été projetée pour inclure des cas avec toutes les vues mammographiques disponibles, les résultats radiologiques, le diagnostic prouvé par biopsie, les antécédents cliniques des patientes et des informations concernant le style de vie des patientes. Chaque examen de chaque cas comprend quatre vues (deux vues de chaque sein : cranio-caudale, et médio-latérale oblique ou MLO). Pour aborder les aspects pédagogiques et de recherche, la base de données relie chaque mammographie au contour du sein, à la limite du muscle pectoral (vues MLO uniquement), aux contours des masses (si présentes), les régions des groupes de calcifications et le nombre de calcifications (le cas échéant), ainsi que les emplacements et les détails de toutes les autres caractéristiques d'intérêt. La base de données mammographiques prend également en charge l'inclusion de plusieurs examens mammographiques du même patient effectués à différents instants de temps.

II.6.3 La base de données IRMA

IRMA (**Image Retrieval in Medical Applications**) est un projet coopératif du Département de radiologie de diagnostic, du Département d'informatique médicale, de la Division de traitement d'images médicales et de l'Université de technologie (**RWTH Aachen**) en Allemagne. L'objectif du projet est le développement et la mise en œuvre de méthodes de haut niveau pour la recherche d'images basée sur le contenu avec une application prototypique à des tâches médico-diagnostiques sur une archive d'images radiologiques.

II.6.4 La base de données DDSM

La base de données numérique pour le dépistage de la mammographie (**DDSM Digital Database for Screening Mammography**) est une ressource destinée à la communauté de recherche sur l'analyse d'images mammographiques. Le soutien principal de ce projet était une subvention du programme de recherche sur le cancer du sein du Commandement de la recherche médicale et du matériel de l'armée américaine.

Le projet DDSM est un effort de collaboration impliquant :

- ❖ l'hôpital « Massachusetts General Hospital » (**MGH**),
- ❖ l'université « Wake Forest University » (**WFU**)
- ❖ l'hôpital « Washington University of St. Louis, School of Medicine » (**WUSTL**).
- ❖ l'Université de Floride du Sud.
- ❖ le laboratoire de recherches « **Sandia National Laboratories** » [44]

La base de données est divisée en 43 volumes, et chaque volume est divisé en un certain nombre d'études : études des cas normaux, études des cas contenant des anomalies bénignes et études contenant des anomalies malignes. Au total, il y a 2620 cas. Le but de cette base de données est de fournir un large ensemble de mammographies dans un format numérique qui peut être utilisé par les chercheurs pour évaluer et comparer les performances des systèmes d'aide au diagnostic/détection (CAD) du cancer du sein.

II.7 Conclusion

Dans ce chapitre, la notion du cancer du sein ainsi que les outils de l'imagerie médicale permettant le diagnostic de ce type de cancer ont été présentés. Une grande attention a été consacrée à l'étude des spécifications des pathologies mammaires à savoir les masses et les micro-calcifications. Une telle étude est fortement associée à la présentation des standards adoptés par les radiologues pour classifier les lésions mammaires. L'objectif de détailler ces études est de mieux introduire le chapitre suivant qui sera consacré à l'étude des outils et méthodes utilisés dans ce travail, notamment la base de données mammographique DDSM, et la classification des anomalies en se basant sur les descripteurs d'Haralick présentés au premier chapitre.

CHAPITRE III

ANALYSE DES BESOINS

ET CONCEPTION

III.1 Introduction

Ce chapitre est consacré à la description des outils exploités dans ce travail, ainsi que la préparation des données et la conception du modèle d'apprentissage, afin d'analyser l'impact du choix de la distance « d » sur le calcul des descripteurs de Haralick et sur les résultats de classification.

En ce qui concerne les données, nous utiliserons la base de données numérique pour la mammographie de dépistage reconnue en anglais sous le nom (**DDSM : Digital Database for Screening Mammography**), cette base a été rassemblée par un groupe de chercheurs de l'Université du sud de la Floride et a été largement utilisée par la communauté scientifique dans le domaine du cancer du sein.

Pour la classification, nous utiliserons un réseau de neurones artificiel multi couches, construit avec les bibliothèques TensorFlow et Keras. TensorFlow est un Framework d'apprentissage automatique développé par l'équipe Google Brain, il est devenu en un temps record l'un des frameworks de référence pour le Deep Learning et l'Intelligence Artificielle, utilisé aussi bien dans la recherche qu'en entreprise pour des applications en production.

III.2 Démarche générale adoptée dans ce travail

La représentation de la démarche générale que nous proposons dans ce travail pour l'étude des paramètres de Haralick pour la classification des anomalies mammographiques est illustrée par la **figure III.1**. Dans cette démarche, nous utiliserons les images mammographiques ayant des anomalies et les régions d'intérêts (ROIs) de chaque anomalie seront séparées ; ces ROIs seront utilisés pour la construction des matrices de cooccurrences GLCM et le calcul des 14 descripteurs de Haralick selon plusieurs distances d , les descripteurs ainsi obtenu seront utilisés en entrée du réseau de neurone pour entraîner le modèle de classification des anomalies selon les classes de BI-RADS. Enfin les résultats (taux de reconnaissance) des différentes expérimentations seront comparés pour obtenir la distance d qui donnera la meilleure classification des quatre classe BI-RADS.

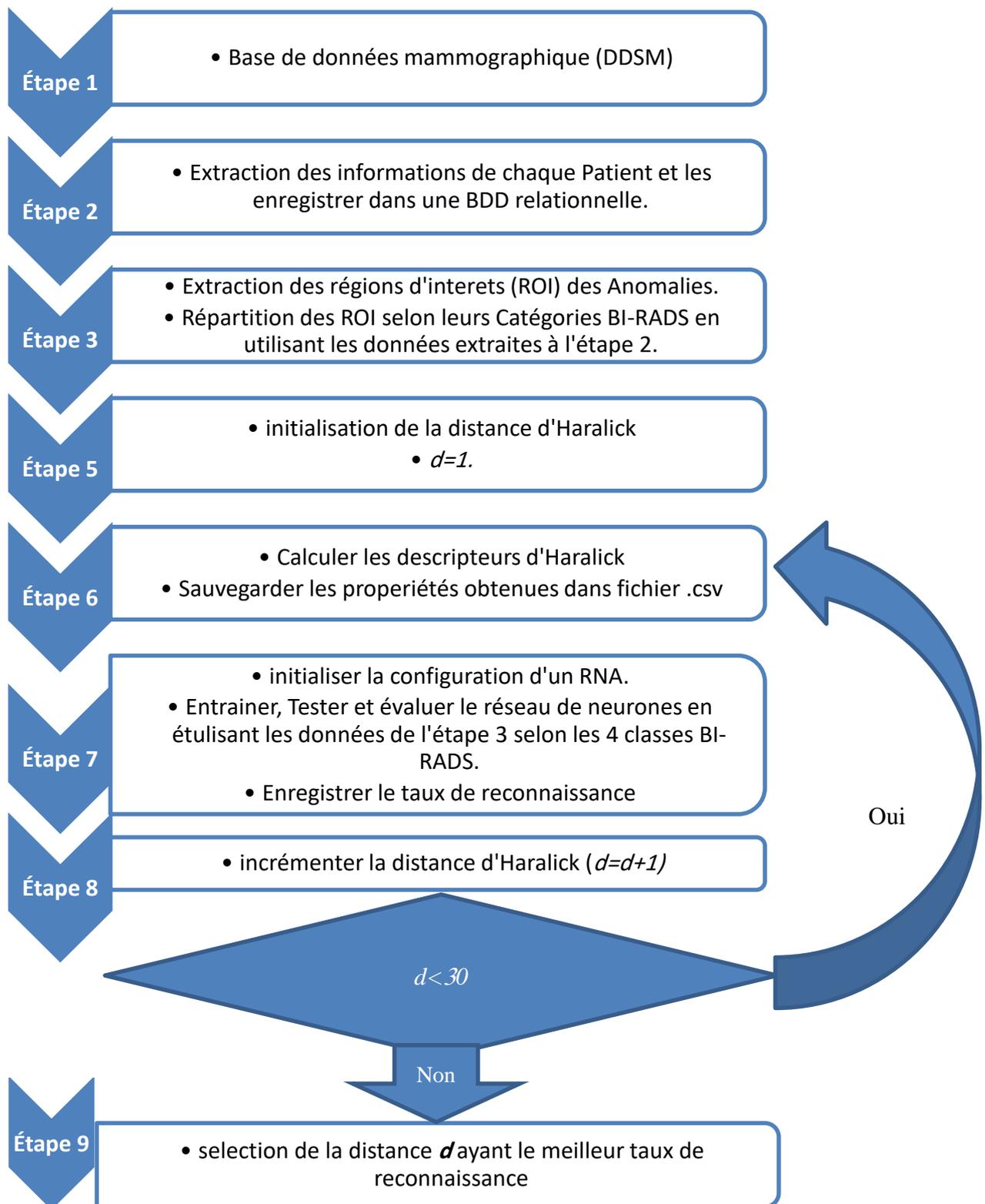


Figure III.1 : Schéma de la démarche générale adoptée pour analyser l'utilisation des paramètres de Haralick dans la classification des anomalies mammographiques.

III.3 La base de données DDSM

Vu sa richesse en données et informations, sa disponibilité, et sa facilité d'utilisation, nous avons choisi la base de données DDSM pour la réalisation de ce travail. Cette base de données a été largement utilisée par la communauté de recherche. Elle est conservée à l'Université de Floride du Sud dans le but de la maintenir accessible sur le Web.

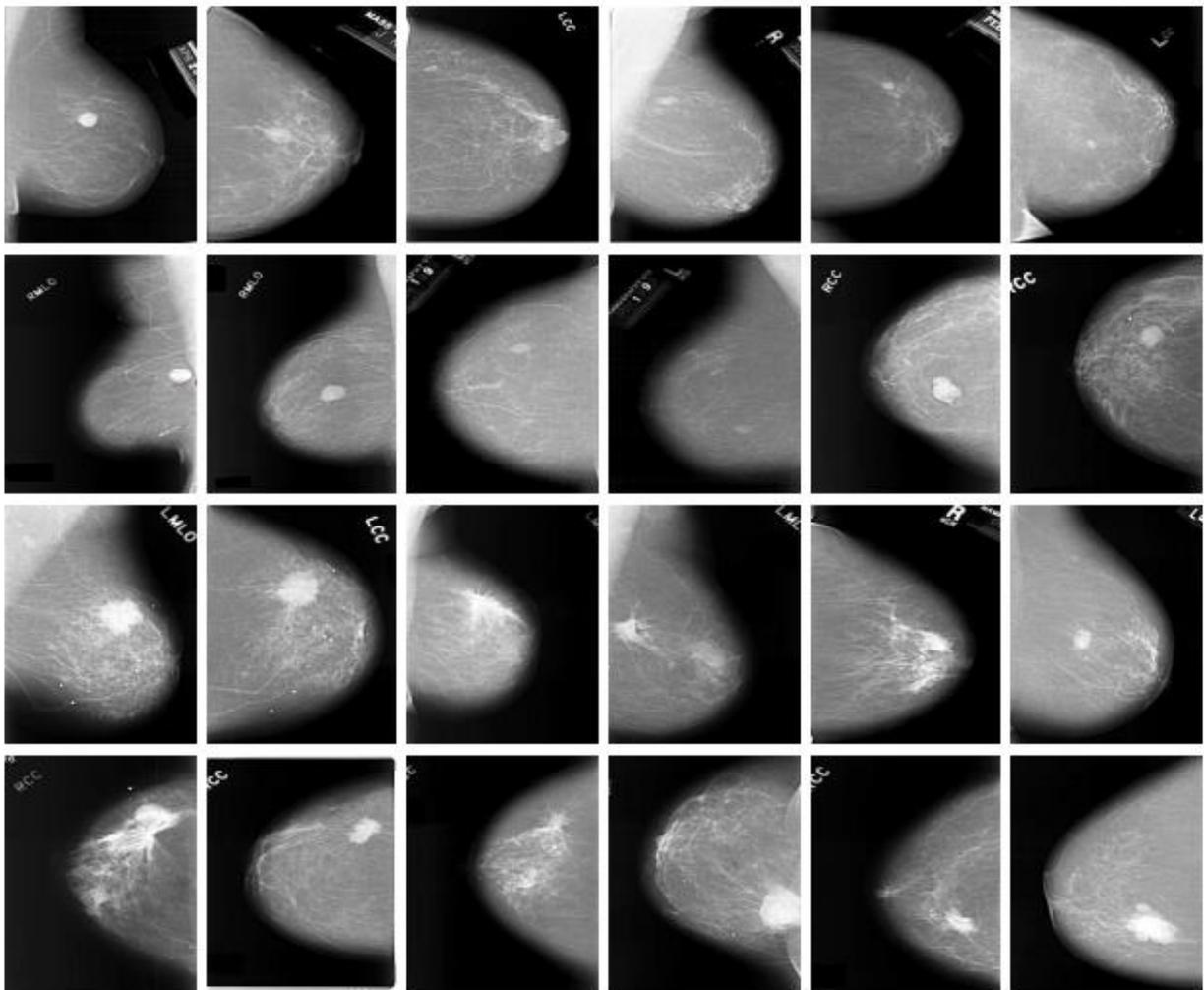


Figure III.2 : quelques échantillons de la base DDSM.

Les images mammographiques ont été numérisées de différentes manières. A l'hôpital « MGH », deux types de scanners ont été utilisés : le « **DBA** » et le « **Howtek** ». Pour ces deux scanners, les niveaux de gris sont codés à 16 et 12 bits respectivement et leur résolution est de 42 et 43.5 microns. A l'université « **WFU** », un scanner « **lumisys** » avec niveaux de gris à 12 bits et à résolution de 50 microns a été utilisé. Finalement, à l'hôpital « **WUSTL** », un scanner « **Howtek** » de mêmes caractéristiques que celui utilisé à « l'MGH » est adopté [52].

Site de numérisation	Type de scanner	Codage niveau de gris	Résolution en micron
MGH	DBA	16	42
MGH	Howtek	12	43.5
WFU	lumisys	12	50
WUSTL	Howtek	12	43.5

Tableau III.1 : Les différents sites et scanners utilisés dans la numérisation de la base DDSM.

La base de données est divisée en 43 volumes : 12 normale (sans anomalies), 15 cancer (lésion malignes), 14 bénignes (des anomalies bénignes), et 2 bénins sans rappel.

Type de lésions	Nombres de volumes	Nombres total des cas	Volume en GB
Normale	12	695	55.7
Maligne	15	914	81.8
Bénigne	14	870	81.1
Bénigne sans rappel	02	141	12.3
Total	43	2620	230.9

Tableau III.2 : Répartition des cas selon le type de pathologie Dans la base DDSM.

Chaque cas d'étude est stocké dans un répertoire distinct, ce répertoire contient 6 à 10 fichiers, la figure III.3 illustre le contenu d'un répertoire d'un cas d'étude.

> Cancer 01 > case0029

Name	Date modified	Type	Size
 TAPE_C_0029_1.COMB.16_PGM	05-Mar-00 17:00	16_PGM File	454 KB
 C-0029-1.ics	05-Mar-00 17:00	Fichier iCalendar	1 KB
 C_0029_1.LEFT_CC.LJPEG	05-Mar-00 17:00	LJPEG File	16,083 KB
 C_0029_1.LEFT_MLO.LJPEG	05-Mar-00 17:00	LJPEG File	16,516 KB
 C_0029_1.RIGHT_CC.LJPEG	05-Mar-00 17:00	LJPEG File	17,386 KB
 C_0029_1.RIGHT_MLO.LJPEG	05-Mar-00 17:00	LJPEG File	18,967 KB
 C_0029_1.LEFT_CC.OVERLAY	05-Mar-00 17:00	OVERLAY File	5 KB
 C_0029_1.LEFT_MLO.OVERLAY	05-Mar-00 17:00	OVERLAY File	6 KB

Figure III.3: La liste des fichiers du cas C-0029 volume : Cancer N° 1.

- ❖ Les 4 fichiers « **.LJPEG** » correspondent aux quatre images représentant les incidences CC et MLO de chaque sein (deux par sein), ces images sont compressées au format « **LJPEG** » un format sans perte d'information. Les fichiers « **.LJPEG** » sont compressés à l'aide d'un logiciel de compression « **JPEG Lossless** » développé à l'Université de Stanford. Ce logiciel est défectueux. Une fois qu'une image est compressée en l'utilisant, elle ne peut être décompressée qu'en utilisant le même logiciel ; lorsque chaque image était compressée, les paramètres d'entrée de ligne et de colonne du compresseur étaient systématiquement inversés. Cela signifie qu'une fois que vous décompressez une image DDSM avec le logiciel Stanford, il rapportera les lignes et les colonnes de l'image inversées. Cependant, les données d'image contenues dans les fichiers « **.ics** » sont correctes [52].
- ❖ Un fichier « **.ics** » qui est un fichier en format ASCII. Il contient les informations importantes du cas telles que l'âge du patient, la date d'étude et de numérisation, le type et les caractéristiques de la lésion si elle existe. Les conditions de numérisation et les informations de la mammographie (le nom du cliché « **.LJPEG** » son nombre de lignes et de colonnes, codage en niveau de gris, la résolution en micron, et éventuellement l'existence de fichier « **.OVERLAY** » qui possède les informations de l'anomalie, la figure III.4 représente un exemple d'un fichier « **.ics** ».

```

ics_version 1.0
filename C-0029-1
DATE_OF_STUDY 2 3 1993
PATIENT_AGE 66
FILM
FILM_TYPE REGULAR
DENSITY 3
DATE_DIGITIZED 9 10 1997
DIGITIZER LUMISYS LASER
SEQUENCE
LEFT_CC LINES 4648 PIXELS_PER_LINE 2672 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
LEFT_MLO LINES 4680 PIXELS_PER_LINE 2664 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
RIGHT_CC LINES 4704 PIXELS_PER_LINE 2816 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY
RIGHT_MLO LINES 4648 PIXELS_PER_LINE 2864 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY

```

Figure III.4 : Exemple d'un fichier « **.ics** ».

- ❖ un fichier « **.OVERLAY** » affecté à chaque incidence ayant une anomalie, le fichier contient les informations de l'anomalie tel que le nombre de lésions dans l'image, leurs

types (masses/Calcification), la classification suivant BI-RADS, la subtilité, la pathologie (maligne/Bénigne), et la délimitation des contours des anomalies existantes.

```

TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC-FINE_LINEAR_BRANCHING
DISTRIBUTION REGIONAL
ASSESSMENT 5
SUBTLETY 4
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 4
BOUNDARY
8 1368 4 4 4 4 4 4 4 4 2 2 2 2 2 2 2 2 2 2 ... 0 0 0 0 0 0 0 0 0 0 1 #
    CORE
168 1824 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 ... 1 0 1 1 0 1 1 0 1 1 #
    CORE
384 1848 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 ... 0 0 0 0 0 0 0 0 0 0 0 #
    CORE
368 2192 6 6 6 6 6 6 6 6 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 #
    
```

Figure III.5 : Exemple d'un fichier « .OVERLAY ».

Les contours des anomalies sont définis par une séquence de chiffres (de 0 à 7) et séparés par des espaces et terminés par le caractère spécial « # », les deux premiers blocs consécutifs représentent les coordonnées (x, y) du premier pixel, les chiffres suivantes défini un déplacement d'un pixel dans une direction donnée pour former les contours, la codification de déplacement après chaque chiffre est illustré par la figure III.6.

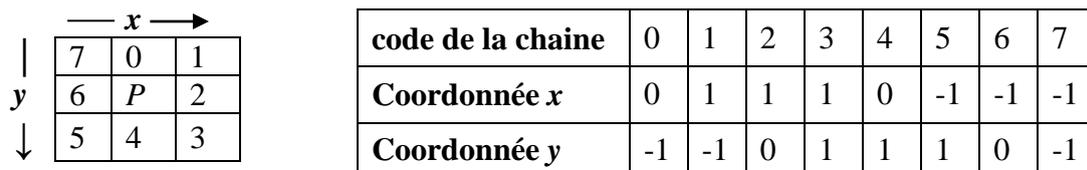


Figure III.6 : Codification de déplacement dans la chaine des contours d'un fichier « .OVERLAY ».

- ❖ Et enfin le fichier « .16_PGM » est une image qui englobe les quatre mammographies (LGPEG) sous-échantillonnées et concaténées. Ils sont stockés dans un format PGM (Portable Gray Map) et codés sur 16 bits pour chaque Pixel. Un exemple du résultat de l'affichage d'un fichier « *.16_PGM » est présenté sur la Figure III.7. Le but de ce type de

fichiers est simplement de fournir une vue globale et rapide sur les quatre images mammographiques LJPEG.

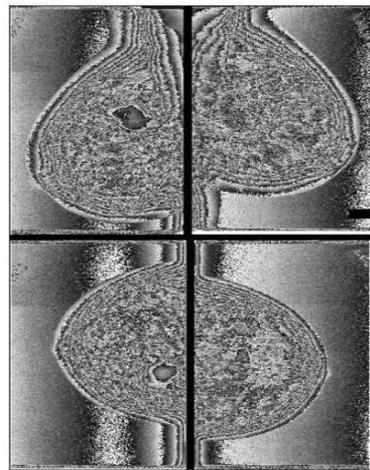


Figure III.7 : Exemple d'un fichier ".16_PGM" représentant les quatre mammographies sous-échantillonnées.

III.4 Extraction des images de la base et préparation de données

Vu à la richesse en données et informations de la base de données mammographique DDSM présenté à la section précédente, et sa disponibilité, nous l'avons choisi pour la réalisation de ce travail.

Les images de cette base de données étant répartie sur plusieurs volumes et répertoires, et les informations des différents cas sont stockés dans des fichiers ASCII, d'où une nécessité d'organiser les données et les informations de cette base de données pour faciliter l'accès et la manipulation des images. Pour ce faire, nous avons opté pour la création d'un système qui extrait les informations à partir de ces fichiers ASCII et en les rangeant dans une base de données relationnelle et qui organisera les informations de cette base de données DDSM. Le model conceptuel de donnés **MCD** de cette base est représenté par la **figure III.8**.

III.4.1 Description des tables de la base de donnés

1. **La table IcsInfos :** Dans cette table sont rangés les informations générales de chaque patient, l'âge, la date de dépistage, la date de numérisation des mammographies, le degré de densité de la glande mammaire selon la codification « **ACR** » ; type du scanner utilisé, type du film et l'**URI** du répertoire qui contient les fichiers du patient.

2. **La table ImageInfos** : Contient les informations de chaque mammographie de la base de données DDSM, **fileName** : le nom du fichier image ; **h** : la hauteur de l'image en pixel ; **w** : son largeur; **anomalieExist** : un booléenne indiquant l'existence ou non des anomalies dans l'image ; **nbParBit** : le nombre de bit sur lequel le niveau de gris est codé (12 ou 16 bit) ; **resolution** : la résolution en micron du scanneur utilisé pour la numérisation.

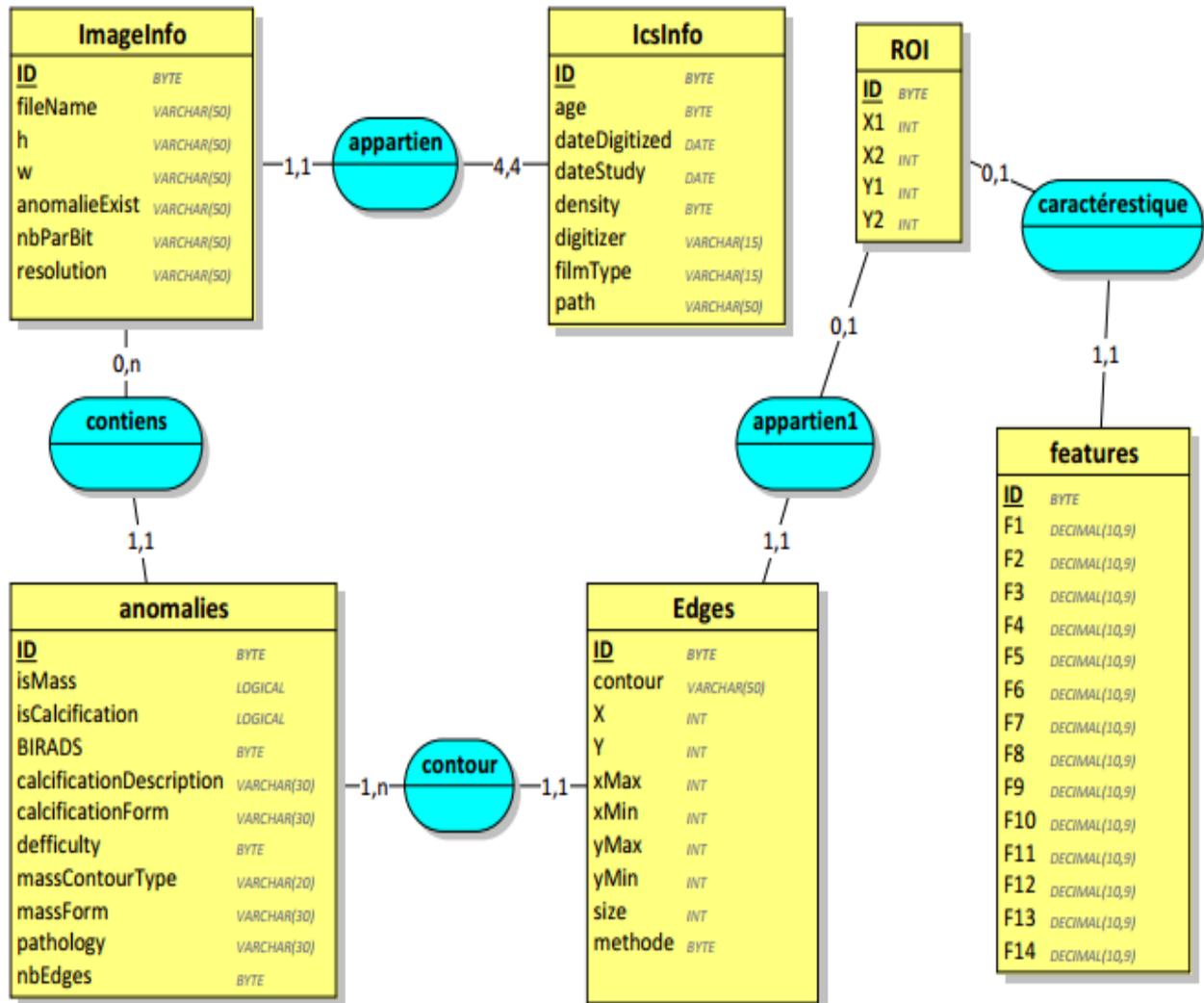


Figure III.8 : Model conceptuel de données pour la base DDSM.

3. **La table anomalies** : Contient les informations des anomalies détectées sur chaque mammographie. **isMass**, **isCalcification** : deux booléenne indiquant s'il s'agit d'une masse ou d'une calcification, **BIRADS** : un entier indiquant la classe BI-RADS de l'anomalie, **calcificationDescription** : une description de la calcification si l'anomalie est de type calcification, **calcificationForme** : la forme de la calcification, **massContourType** : type des

contours de la masse si l'anomalie est de type masse, **massForm** : la forme de la masse, **pathology** : la pathologie de l'anomalie approuvée après une biopsie (bénigne ou maligne), **nbEdges** : nombre de contours détectés dans cette images.

4. **La table Edges** : Contient les informations des contours de chaque lésion ; **contour** : est une chaîne codée qui représente les contours de l'anomalie tel qu'elle est codifiée dans le fichier « .OVERLAY » ; **X** : l'abscisse du point de départ du contour ; **Y** : l'ordonnée du point de départ du contour ; **xMax, xMin, yMax, yMin** : sont les valeurs maximale et minimale pris par le tracé des contours par rapport au deux axes, horizontale et verticale.

5. **La table ROI (Region of interest)** : Contient les coordonnées des régions d'intérêts à découper pour calculer le GLCM et les paramètres de Haralick. il est vrai que, selon les règles des formes normales et le principe des dépendances fonctionnelles, les attributs de cette table sont tous fonctionnellement dépendant de la table « **Edges** », et par conséquent doivent être inclus dans cette dernière ; mais nous avons choisi de créer la table ROI pour simplifier l'extraction des attributs de texture de Haralick et d'optimiser le temps d'accès à la base de données.

6. **La table features** : Cette table sert à stocker les 14 descripteurs de Haralick une fois calculées.

III.4.2 Implémentation de la base de données

L'implémentation de cette base de données et de ce système est pour faciliter la manipulation des données est réalisée en utilisant le langage Java, et l'IDE Eclipse, « l'SGBD utiliser est « Apache Derby » qui est une base de données relationnelle, supportée par « l'Apache Software Foundation », publiée en open source, et libre de droits. L'application est entièrement écrite en Java, ce qui la rend portable sur toutes les machines possédant une JVM.

Derby a la particularité d'être une base embarquée : il est possible de lancer le moteur de Derby dans la même JVM que son application. L'accès aux données se fait via la technologie JPA (**Java Persistence API**), JPA offre un modèle d'ORM (**Object Relational Mapping**) qui assure l'indépendance des traitements aux services de persistance (Derby en l'occurrence).

Cette application est utilisée pour lire automatiquement les informations contenues dans

les fichiers texte « **.ics** » et « **.OVERLAY** » et enrichir la base de données par ses informations, par la suite ces informations seront utilisées pour la préparation de la base d'expérimentations et l'extraction des caractéristiques utilisées dans la phase d'expérimentations.

III.4.3 Une application JAVA pour extraire les sous images de régions d'intérêt

Afin de faciliter l'exploitation de la base des images DDSM, nous avons opté pour la création d'une application en JAVA, son but principal est d'extraire les informations stockés dans des fichiers ASCII (.ics et .OVERLAY) et les organisent dans une base de données relationnelle, de cette façon nous auront une carte détaillé et bien organisé de la base mammographique DDSM. Cette carte ainsi obtenu, sera utilisé pour extraire à partir chaque mammographie de la base DDSM une sous-image de la région d'intérêt (ROI) qui englobe l'anomalie, cette ROI est placé dans un sous répertoire qui porte le nom de la classe BI-RADS de l'anomalie.

Les régions d'intérêts organisées dans les sous répertoires de chaque classe BI-RADS seront utilisés par la suite par un autre module de l'application afin de calculer les 14 descripteurs de Haralick, et qui stock ces derniers dans la table « **features** » de la base de données relationnelle.

III.4.4 Choix de la méthode de classification des anomalies en BI-RADS

L'objectif de notre travail étant l'analyse de l'impact du choix de la distance dans le calcul des paramètres de Haralick sur les résultats de classification des anomalies mammaires, alors le choix de la méthode de classification a une importance primordiale.

Dans notre cas, les données d'apprentissages sont étiquetées, donc l'apprentissage approprié est l'apprentissage supervisé.

Il existe plusieurs méthodes d'apprentissage supervisé. Pour connaitre qu'elle est la meilleur méthode de classification à utiliser dans ce cas, nous avons conçu un programme en Python pour comparer et mettre à l'épreuve quelques classifieurs les plus fréquents, il s'agit des méthodes suivante :

- SVM : « Support Vector Machine » Machines à vecteurs de support, appelés aussi séparateurs à vaste marge, sont des techniques d'apprentissage supervisées destinées à résoudre des problèmes de discrimination et de régression. Cette méthode repose sur des bases mathématiques solides, elle s'articule sur la théorie de l'apprentissage statistique et sur la théorie des bornes de Vapnik et Chervonenkis [53]. Le principe de classification est de définir l'hyperplan qui sépare les classes tout en maximisant la marge qui les sépare. la sélection de vecteurs supports qui représentent les vecteurs discriminant grâce auxquels est déterminée l'hyperplan. ces vecteurs supports sont utilisés pour classer les nouveaux cas,
- RL: « Logistic regression » Régression logistique : C'est une technique permettant de trouver une fonction de prédiction d'un variable dit variable expliquée dichotomique à partir des données statistiques dite variable explicatives. Il s'agit en fait de connaître les facteurs associés à un phénomène en élaborant un modèle de prédiction. La popularité de cette méthode est bien connue dans les sciences de la santé et en sciences humaines, où la variable à prédire est la présence ou l'absence d'une maladie ou autre phénomènes, et on veut étudier l'influence des facteurs observés sur celui-ci.
- LDA : « Linear Discriminant Analysis » : L'analyse discriminante linéaire est utilisée comme outil de classification, de réduction de dimension et de visualisation des données. Malgré sa simplicité, LDA produit souvent des résultats de classification robustes, décentés et interprétables. Lorsqu'il s'agit de résoudre des problèmes de classification dans le monde réel, la LDA est souvent la première méthode d'analyse comparative avant d'utiliser d'autres méthodes plus compliquées et flexibles.
- KNN : « K Neighbors Classifier » K plus proches voisins : L'algorithme des K-plus proches voisins (KNN) utilise la similarité des caractéristiques pour prédire les classes des nouvelles données, ce qui signifie que le nouveau données se verra attribuer une classe en fonction de son degré de correspondance avec les données de l'ensemble d'apprentissage..
- DTC : « Decision Tree Classifier » Arbre de décision : C'est un outil d'aide à la décision ou d'exploration de données, permet de représenter un ensemble de choix sous la forme graphique d'un arbre. C'est une des méthodes d'apprentissage supervisé les plus populaires pour les problèmes de classification de données. Concrètement, un arbre de décision modélise une hiérarchie de tests pour prédire un résultat.

- NB : « Naive Bayes » Classifieur Bayésien naïve : Les algorithmes naïfs de Bayes sont une technique de classification basée sur l'application du théorème de Bayes avec une forte hypothèse que tous les prédicteurs sont indépendants les uns des autres. En termes simples, l'hypothèse est que la présence d'une caractéristique dans une classe est indépendante de la présence de toute autre caractéristique de la même classe.
- RNA : « Artificial Neural Network » réseau de neurones artificiels :

Les 6 premiers classifieurs ont été implémentés en utilisant la bibliothèque d'apprentissage automatique «scikit-learn», pour la réalisation de ces tests, on a sélectionné 2400 images, 600 de chaque la classe. Ces images sont divisées en deux partis, 80 % pour l'entraînement et 20 % pour les tests. la figure III.9 illustre le résultat de comparaison.

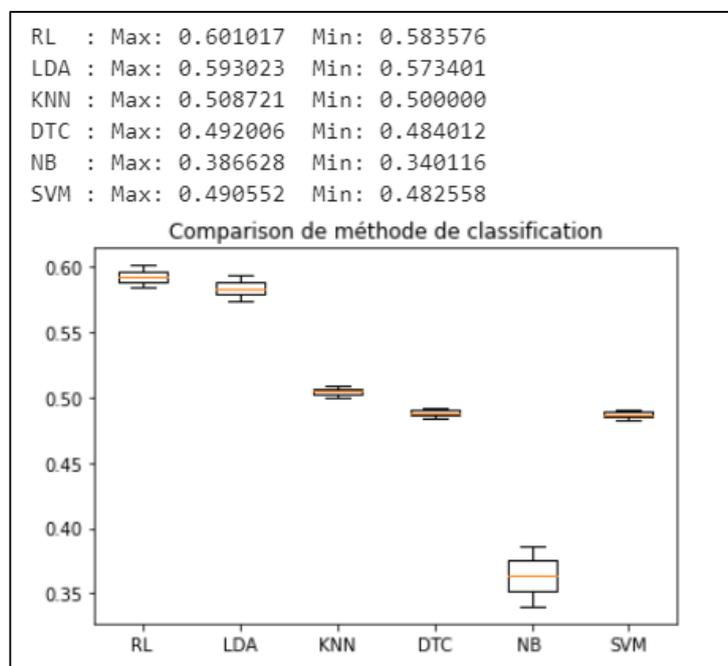


Figure III.9 Résultats de comparaison entre les 6 méthodes de classification.

Pour le réseau de neurones artificiel **RNA**, nous avons utilisé un modèle conçu avec la bibliothèque Keras et TensorFlow en Backend, et on l'entraîne avec la même base de données préparée à la section précédente. La configuration de ce réseau de neurones a été obtenue suite à des essais expérimentaux, il est configuré comme suit :

- Nombre de neurones de la couche d'entrée : 14.
- Nombre de couches cachées : 2.
- Nombre de neurones dans les couches cachées : 26.

- Nombre de neurones dans la couche de sortie : 4.
- Fonction d'activation de la couche de sortie : « sigmoid ».
- Fonction d'activation des autres couches : « relu »

Les résultats obtenus par le RNA sont présenté à la **figure III.10**

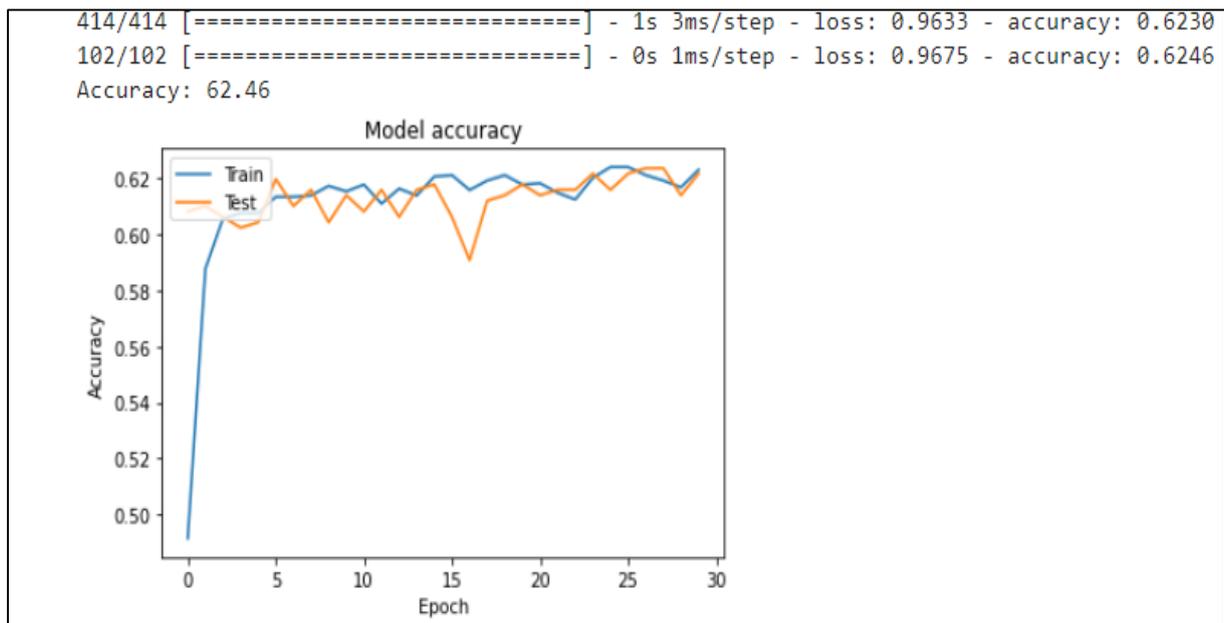


Figure III.10 Résultats de classification par un réseau de neurones artificiel.

A partir de ces résultats, nous remarquons que seule la précision obtenu par les méthodes RL, LDA et RNA sont au-dessus de 50 %, pour les autres classifieurs la précision est entre 34.01% et 50.87%, donc les techniques de classification intéressantes sont la régression logistique, l'analyse discriminante linéaire, et le réseau de neurones artificiels, et ce dernier a marqué un peu plus de précision.

Sur la lumière de ces résultats, nous avons choisi le réseau de neurones artificiels RNA comme classifieur à utiliser pour les expérimentations qui seront effectués dans ce travail.

III.5 Modèle utilisé dans ce travail

Il existe une multitude de réseaux de neurones, variés par leur architecture, leur méthode d'apprentissage et beaucoup d'autres caractéristiques. Pour chaque application il faut choisir le modèle qui lui est adapté. Ce choix est influencé par plusieurs facteurs, les plus significatives sont :

- Le type de la fonction à modéliser (classification, clustering, diagnostic ou reconnaissance).

- Nature et disposition des données d'apprentissage (statique, dynamique, aléatoire, d'un flux contenu...etc.)
- Ressources matériels et logiciels disponible pour l'implémentation et l'apprentissage du réseau.

Dans ce travail, nous cherchons à analyser les 14 paramètres de Haralick extraites des images mammographiques, donc il s'agit d'une classification supervisée, un perceptron multicouche sera un choix opportun.

Notre réseau de neurones est conçu à l'aide de la bibliothèque Keras avec Tensorflow en backend. Nous avons utilisé le modèle séquentiel avec des couches totalement connecté (Dense), ce model se compose de 4 couche : la première couche pour les entrés (input layer), elle se compose de 14 neurones réservé aux 14 descripteurs de Haralick.

En suite deux autres couche cachées (hidden layers). Il n'existe pas de règle générale pour déterminer le nombre de neurones par les couches cachées, cependant des expériences empiriques ont démontré l'efficacité de certaines choix en dépendance avec le nombre des neurones d'entrés et/ou le nombre de neurone de sortie [52]. Dans ce travail nous avons choisi 26 neurones pour chaque couche cachée, Ce choix a été fait après plusieurs expériences sur plusieurs modèles, et on a gardé celui qui donne la meilleure précision. Notons que chaque neurone supplémentaire permet de prendre en compte des profils spécifiques des neurones d'entrée. Un nombre plus important permet donc de mieux coller aux données présentées mais diminue la capacité de généralisation du réseau.

Enfin, pour la couche de sortie, le nombre de neurones suit le nombre de classe de sortie, dans notre cas, on veut classer les anomalies suivant la classification BI-RADS ; ce dernier à sept classes (voir section **II.5.3**). Les classes 0 et 1 concernent les mammographies sans anomalies, donc ne elles ne sont pas pris en compte par notre classifieur. La 6eme classe est écartée car elle représente les cas dont les résultats sont approuvés après une biopsie. Donc seulement les quatre classes restantes (2, 3, 4 et 5) seront considérées, Ainsi le nombre de neurones de la couche de sortie est fixé à 4 neurones. **La figure III.17** illustre le schéma général du modèle à utiliser.

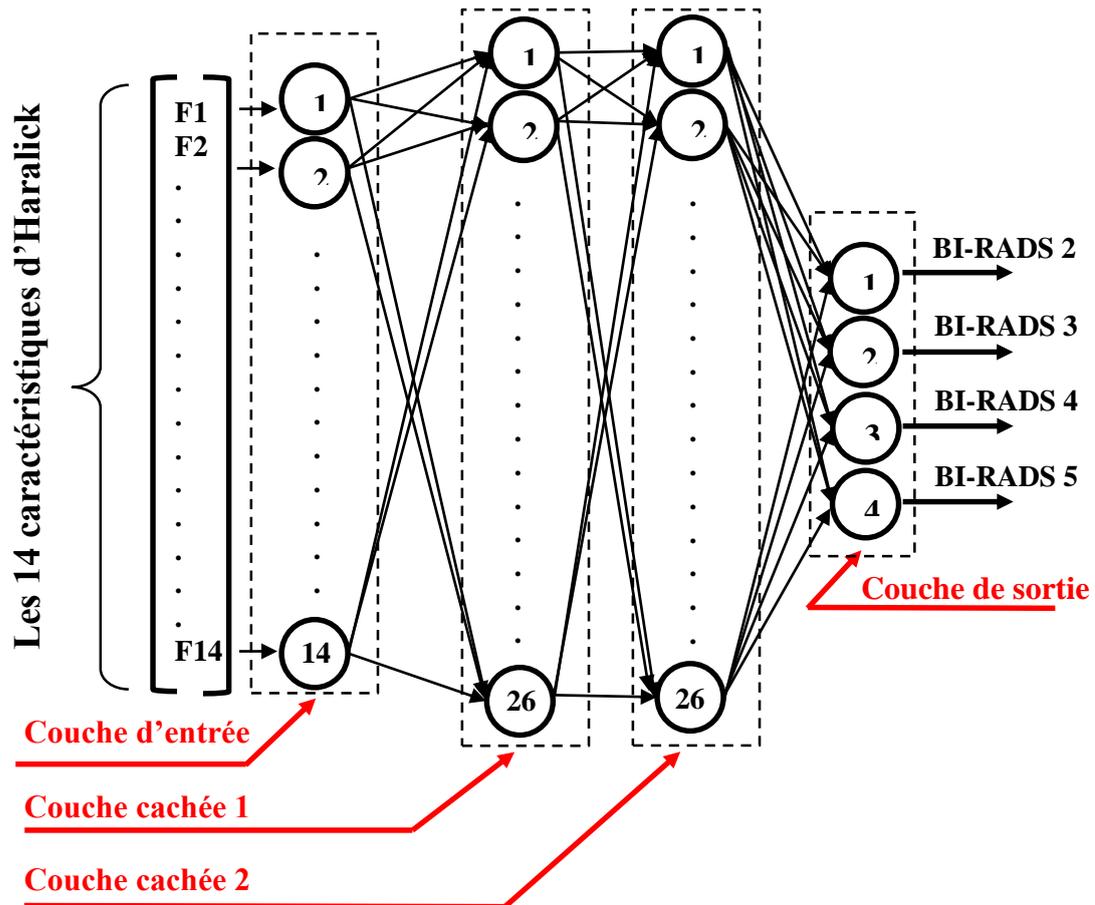


Figure III.11 *Modèle du réseau de neurone utilisé pour la classification des anomalies mammaires.*

III.6 Conclusion

Une description générale des méthodes et des outils utilisés pour la réalisation de ce travail ont été détaillés dans ce chapitre. Commenant par l'établissement des étapes à suivre dans la démarche générale de ce travail. Ensuite l'étude des propriétés de la base de données mammographiques et la conception une application en JAVA pour organiser les informations de la base de données dans une base relationnel qui nous permettra par la suite de faciliter l'extraction des régions d'intérêts qui concernent les anomalies de chaque classe BI-RADS et les sous images correspondantes. Nous avons justifié aussi le choix de la technique de classification en se basant sur un algorithme de comparaison entre les méthodes les plus utilisés.

Puis, nous avons présenté la conception du modèle à utiliser dans le chapitre suivant pour la classification et l'analyse des descripteurs de Haralick.

CHAPITRE IV

Implémentation et résultats expérimentaux

IV.1 Introduction

Dans ce chapitre nous allons présenter notre solution pour analyser les paramètres de Haralick, commençant par l'environnement matériel et logiciel utilisé dans les différentes étapes de ce travail, Ensuite les résultats des expériences seront comparés pour déterminer la meilleur distance utilisé dans le calcul du matrice de cooccurrence pour avoir la bonne précision, Enfin nous donnons une discussion sur les résultats obtenus.

IV.2 Environnement de développement

Comme on a vu dans le chapitre précédent, notre travail se divise en deux parties principales : la première concerne la préparation et l'extraction des régions d'intérêts (ROIs) de la base de données mammographique DDSM et de calculer les caractéristiques de Haralick à partir de ces ROIs.

Tandis que la deuxième concerne la création et l'entraînement d'un modèle de classification de ces ROIs en trois classes suivant la classification BI-RADS de ACR à l'aide d'un classifieur réalisé par un réseau de neurones (RNA) pour déterminer la meilleur distance à utiliser dans le calcul du matrice de cooccurrence qui nous conduira au meilleur taux de reconnaissance.

IV.2.1 Environnement pour la préparation des données

Pour la première partie de préparation des donnés, nous avons utilisé deux langages de programmation dans deux environnement différents, ces deux langages sont :

- Python sous le système d'exploitation Linux utilisé pour les prétraitements des images, c.à.d. la décompression des fichiers .LJPEG, la réduction du nombre de niveau de gris des images qui sont codé sur 16 ou 12 bit pour avoir des image standard (en niveaux de gris) codé sur 8 bit. Cette réduction engendre une perte de qualité, mais cette perte est justifier par la réduction de la complexité des calculs, car la taille de la matrice de cooccurrence d'une image codée sur 16 bits est une matrice carré de $2^{16} \times 2^{16} = 4.294.967.296$ entrés, dont la difficulté de calcule devient très élevé.
- Le langage de programmation JAVA sous l'environnement de développement Eclipse pour la réalisation d'une application utilisée pour l'extraction et l'organisation des informations contenu dans les fichiers ASCII de la base mammographique DDSM, et pour l'extraction des régions d'intérêt ROIs des anomalies

mammographiques, et qui nous permettra par la suite de calculer les matrices de cooccurrence et les descripteurs de Haralick.

Le matériel utilisé pour ces tâches est un PC portable ayant les caractéristiques suivantes :

- Type : PC ASUS X5415.
- Processeur : Intel Celeron CPU N3060 @ 1.60 GHz.
- Mémoire installée (RAM) : 4.00Go.
- Système d'exploitation : Deux systèmes d'exploitation coexiste sur le PC :
 - Microsoft Windows 10 Professionnel 64bits, installé d'origine.
 - Linux UBUNTU 19.4. Le but de son installation est d'exécuter le programme de décompression des fichiers image .LGPEG développé à l'Université de Stanford (voir la section III.3).

IV.2.1.1 Téléchargement, Décompression, et Normalisation des images

La base de données mammographique DDSM est accessible gratuitement au site web (<http://www.eng.usf.edu/cvprg/Mammography/Database.html>); elle contient 2620 cas réparti sur 43 volumes d'une taille totale de **230.9 GO**, le téléchargement a pris une vingtaine de jours.

Une fois le téléchargement terminé, nous avons écrit un programme Python de deux script le premier « convert.py » qui explore les répertoires des images, et qui fait appel à un autre script « ljpeg.py » (voire les annexes 1 et 2). Ces deux scripts sont exécutés à travers le système d'exploitation « Linux » pour décoder les mammographies « LJPEG » de la base de données DDSM et pour réduire le nombre de niveau de gris de (12 ou 16 bits) à 8 bits, et en même temps pour normaliser les images afin d'améliorer le contraste des images en utilisant la bibliothèque Open CV.

IV.2.1.2 Extraction et organisation des annotations de la base DDSM

Afin de faciliter l'exploitation de la base de données mammographique DDSM, vu que ces informations sont arrangées dans plusieurs fichiers texte, nous avons créé une application JAVA pour l'extraction et l'organisation de ces annotations dans une base de données relationnelle bien structuré. Cette application se compose des packages suivant :

- **BDD** : regroupe les classes de définition des tables de la base de données en respectant les annotations de JPA (Java Persistence API) [53] : L'API Java Persistence est un ensemble de classes et de méthodes permettant de stocker de manière permanente les vastes quantités de données dans une base de données fournie par Oracle Corporation. Il existe actuellement trois implémentations principales de JPA :

- EclipseLink, l'implémentation de référence ;
- Hibernate, projet porté par JBoss, appartenant à RedHat ;
- OpenJPA, projet de la fondation Apache.

Dans notre cas nous avons utilisé **EclipseLink** qui s'intègre parfaitement et d'une manière transparente à Eclipse. La base de données est implémentée en utilisant le SGBD « **Apache Derby** » : un sous - projet Apache DB, est une base de données relationnelle open source entièrement implémentée en Java et disponible sous la licence Apache et qui fournit un pilote JDBC intégré qui permet d'intégrer Derby dans n'importe quelle solution Java [69].

- **Stockage** : un package qui regroupe les classes d'interaction avec la base de données et assure la persistance des objets, ces classes sont :
 - **JPAFactory**: cette classe est responsable de l'accès transactionnel à la base de données en créant une instance d'EntityManager utilisé pour gérer les opérations sur la base de données.
 - **IcsLoader** : cette classe encapsule les méthodes pour parcourir les répertoires de la base de données mammographique, récupérer les annotations des fichiers texte .ics et .OVERLAY les adapter pour les stocker dans la base de données relationnel.
 - **CreateDatabase**: utilise IcsLoader et JPAFactory pour créer la base de données et l'enrichir par les informations extraites des fichiers .ics et .OVERLAY.
- **Features** : regroupe les classes :
 - **GenerateFeatures**: cette classe regroupe les méthodes responsables de la préparation des images par l'extraction des sous images des régions d'intérêt, les placés dans des répertoires qui ont le même nom de la classe BI-RADS, et une autre méthode qui calcule les descripteurs de Haralick en utilisant les classes (GreyLevelCooccurrenceMatrix et Haralick) et en fin ce package nous permettra de sauvegardes ces caractéristiques dans la base de données.
 - **GreyLevelCooccurrenceMatrix**: cette classe comme son nom l'indique, est spécialisé dans le calcul de la matrice de cooccurrence.
 - **Haralick** : cette class utilise la classe GreyLevelCooccurrenceMatrix pour calculer la matrice de cooccurrence et les 14 descripteurs de Haralick.

IV.2.2 Outils d'implémentation du réseau de neurones

IV.2.2.1 Dispositifs matériels

Pour l'implémentation du réseau de neurones artificiel, on a opté pour Google Colab ou Colaboratory, c'est un service cloud, offert gratuitement par Google, basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le « cloud ». Il permet également :

1. De développer des applications en Deep Learning en utilisant des bibliothèques Python populaires telles que Keras, TensorFlow, PyTorch et OpenCV.
2. D'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration.
3. De fournir un environnement matériel très puissant avec 12 Go de RAM, et 107 Go de disque de stockage, un GPU ou TPU configurable en option pour les super calculs.
4. La possibilité d'accéder aux données stockées sur Google Drive.

IV.2.2.2 Outils logiciels

Colab utilise Jupyter Notebook avec le langage Python, c'est est un environnement de développement interactif basé sur le Web pour le code et les données, Jupyter Notebook est flexible, c.à.d. l'interface utilisateur est configurée et organisée pour prendre en charge un large éventail de flux de travail en science des données, en informatique scientifique et en apprentissage automatique. Les principales bibliothèques de Python sont préinstallées, avec la possibilité d'installer d'une façon très simple n'importe quelle autre bibliothèque souhaitée. Parmi les bibliothèques préinstallées on trouve :

- **TensorFlow** : TensorFlow est un outil open source d'apprentissage automatique développé par Google. Doté d'une interface pour Python, ce qui lui rend très populaire dans la communauté des chercheurs en machine Learning et Depp Learning et Data sciences. Cette bibliothèque est organisée en couches (voire la **figure IV.1**), L'objectif des différentes couches est décrit dans **le tableau IV.1** [54].

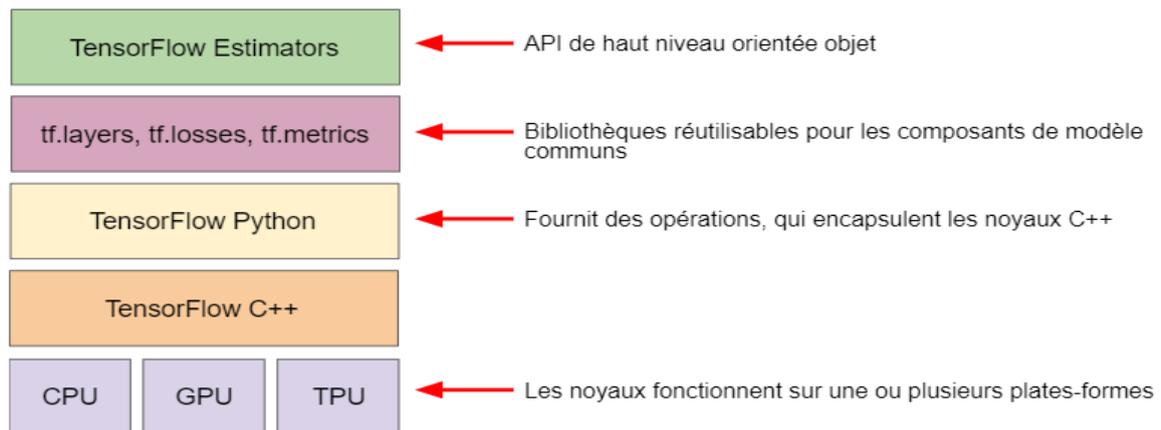


Figure IV.1 : Les couches de la bibliothèque TensorFlow

Kit(s)	Description
Estimator (tf.estimator)	API orientée objet de haut niveau.
tf.layers/tf.losses/tf.metrics	Bibliothèques pour les composants de modèle communs.
TensorFlow	API de niveau inférieur.

Tableau IV.1 : Description des couches de la bibliothèque TensorFlow

Le kit de TensorFlow se compose des deux composants suivants :

- A. Un Protocol Buffer pour le graphe
- B. Un environnement d'exécution qui exécute le graphe (distribué)

Ces deux composants sont analogues au compilateur Java et à la machine virtuelle Java (JVM). À l'instar de la JVM, le kit de TensorFlow est mis en œuvre sur plusieurs plates-formes matérielles : processeurs et GPU.

La bibliothèque TensorFlow est utilisée en « Back end » de notre model RNA.

- **Keras :** Keras est une API d'apprentissage en profondeur écrite en Python, s'exécutant sur la plate-forme d'apprentissage automatique TensorFlow. Il a été développé dans le but de permettre une expérimentation rapide. Être capable de passer de l'idée au résultat le plus rapidement possible. Keras est l'API de haut niveau de TensorFlow, c'est une interface accessible et hautement productive pour résoudre les problèmes d'apprentissage

automatique. Il fournit des abstractions et des éléments de base essentiels pour le développement et l'expédition de solutions d'apprentissage automatique avec une vitesse d'itération élevée. Keras permet de tirer pleinement parti de l'évolutivité et des capacités multiplateformes de TensorFlow : on peut exécuter Keras sur TPU ou sur de grands clusters de GPU, et vous pouvez exporter vos modèles Keras pour les exécuter dans le navigateur ou sur un mobile dispositif [55]. nous avons utilisé cette bibliothèque pour la création du modèle de réseau de neurones.

- **OpenCV (Open Source Computer Vision Library)** C'est une bibliothèque de traitement d'image ou plus généralement une bibliothèque graphique développée principalement avec le langage C++ par la société Intel dont Willow Garage assure le support depuis 2008 [51]. accessible au travers d'API pour les langages C, C++, et Python. Cette bibliothèque est utilisée dans notre travail pour la lecture et l'écriture des images, ainsi que les différents traitements sur les images tel que la conversion en niveau de gris.
- **Numpy** : C'est une bibliothèque numérique apportant le support efficace de larges tableaux multidimensionnels, et de routines mathématiques de haut niveau (fonctions spéciales, algèbre linéaire, statistiques, etc.). [57]
- **Pandas** : est une bibliothèque Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.
- **Scikit-learn** : Scikit-learn (ou Sklearn) est une bibliothèque libre Python dédiée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment par des instituts français d'enseignement supérieur et de recherche comme Inria et Télécom ParisTech. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy [58].
- **Mahotas** : Cette bibliothèque n'est pas installée par défaut dans l'environnement Google colab, Il faut l'installer chaque fois qu'on doit l'utiliser. Tout comme OpenCV, s'appuie sur les tableaux NumPy. Beaucoup de la fonctionnalité mise en œuvre dans Mahotas peuvent être trouvés dans OpenCV, mais dans certains cas, l'interface Mahotas est juste plus facile à utiliser. Nous avons utilisé Mahotas pour calculer les paramètres de Haralick par un script Python (voir l'annexe 3)[5].

Une fois l'exécution de ce script de l'annexe N°3 est terminée, on obtient les 20 fichiers « .csv » contenant chacun les valeurs des descripteurs de Haralick et le label de la classification BI-RADS de chaque Région d'intérêt (ROI). Les valeurs ainsi obtenus sont très variés et dispersés, pour améliorer l'apprentissage on procède à une normalisation des données. Il existe plusieurs méthodes de normalisations, on cite par exemple :

- Zscore: $z = \frac{x - \text{moyen}(x)}{\text{ecartype}(x)}$
- MinMax : $z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$
- Logistic : $z = \frac{1}{1 + e^{-x}}$

Dans ce travail nous avons utilisé la méthode MinMax pour obtenir les valeurs normalisées dans l'intervalle [0, 1].

IV.2.3 Les catégories BI-RADS utilisées

La classification BI-RADS [43] de l'American College of Radiology (ACR) comporte les catégories 3 et 4 qui regroupent les anomalies probablement bénignes pour lesquelles est faite une proposition de surveillance. On retrouve essentiellement dans ces deux catégories des opacités nodulaires et des foyers de microcalcifications ayant une densité similaire. Ces deux classes BI-RADS représentent une catégorie d'anomalies mammographiques à risque moyenne. Distinguer entre les classes BI-RADS 3 et BI-RADS 4 en se basant sur la densité (texture) de l'anomalie reste une tâche difficile et sont certainement les catégories pour laquelle il est plus difficile d'obtenir un consensus entre radiologues [59].

Donc, et pour avoir les meilleurs résultats, nous avons fusionné les données des classes 3 et 4 pour les concéderai comme une seule classe, et cela est logique car les images appartenant à ces deux classes ont la même densité. La liste des classes utilisées par nos modèles sont :

1^{ère} classe : BI-RADS 2

2^{ème} classe : (BI-RADS 3 & 4)

3^{ème} classe : BI-RADS 5

IV.2.4 Les paramètres du classifieur utilisés

Après avoir fait plusieurs expériences sur plusieurs configurations de RNA, nous avons opté pour celui qui donne les meilleurs résultats de classification selon le paramétrage suivant :

- Nombre époques=100 : Le nombre d'époques est un hyper paramètre qui définit le nombre de fois où l'algorithme d'apprentissage fonctionnera sur l'ensemble de données d'apprentissage.

- Batch size (taille du lot) =32 : C'est le nombre d'échantillons traités avant la mise à jour des poids du modèle.
- Nombre de neurones de couche d'entrée : 14.
- Nombre de couches cachées=2.
- Nombre de neurones dans la couches cachées=26.
- Nombre de neurones dans la couche de sortie=3.
- Fonction d'activation= « sigmoid » pour la couche de sortie et « relu » pour les autres.
- Métriques : 'accuracy' : la performance qu'on cherche à optimiser (ici la précision).
- Optimisateur : 'adam' : L'optimisation Adam est une extension de la descent de gradient stochastique, et peut être utilisée à sa place pour mettre à jour plus efficacement les poids du réseau.
- Fonction de perte : « categorical_crossentropy » : C'est de calculer la quantité qu'un modèle doit chercher à minimiser pendant l'entraînement.

Le script de l'annexe N° 4 présente le programme complet en Python qui définit le modèle qui sera utilisé dans l'entraînement et le test et l'affichage des résultats de classification des anomalies selon les trois classes.

IV.3 Expérimentations et résultats

Pour évaluer notre modèle, nous avons sélectionné 2100 régions d'intérêts, 700 de chacune des trois classes, où 50% des anomalies de chaque classe sont des anomalies de type « masse » et la deuxième moitié sont des anomalies de type « calcifications ».

Ces images sont divisées en deux parts, 70 % pour l'entraînement et 30 % pour les tests.

IV.3.1 Critère d'évaluation du modèle

IV.3.1.1 Matrice de confusion

Une matrice de confusion (Confusion Matrix) ou tableau de contingence est un outil permettant de mesurer les performances d'un modèle de Machine Learning en vérifiant notamment à quelle fréquence ses prédictions sont exactes par rapport à la réalité dans des problèmes de classification.

C'est le moyen le plus simple de mesurer les performances d'un problème de classification. Une matrice de confusion n'est rien d'autre qu'un tableau à deux dimensions, à savoir. « Classe Réel » et « classe prédite » et en outre, les deux dimensions ont « Vrais positifs (VP) », « Vrais négatifs (VN) », « Faux positifs (FP) », « Faux négatifs (FN) ». Pour notre cas nous avons trois classes donc la matrice de confusion prend la forme indiqué par la figure IV.6.

		La classe réelle		
		BI-RADS 2	BI-RADS 3&4	BI-RADS 5
La classe prédite	BI-RADS 2	V_2	$F_{3\&4-2}$	F_{5-2}
	BI-RADS 3&4	$F_{2-3\&4}$	$V_{3\&4}$	$F_{5-3\&4}$
	BI-RADS 5	F_{2-5}	$F_{3\&4-5}$	V_5

Figure IV.2 : Matrice de confusion.

Les explications des termes associés à la matrice de confusion présentés dans le tableau IV.2 sont les suivantes :

1. V_x : Nombre des images de la classe X classées correctement dans la classe X.
2. F_{x-y} : Nombre des images de la classe X classées dans la classe Y.

Nous pouvons utiliser la fonction « confusion_matrix » de « sklearn.metrics » pour calculer la matrice de confusion de notre modèle de classification. De la matrice de confusion on peut calculer plusieurs mesures utilisé comme indice de la fiabilité d'un modèle de classification.

IV.3.1.1 Précision de classification (Accuracy)

Il s'agit de la mesure de performance la plus courante pour les algorithmes de classification. Il peut être défini comme le nombre de prédictions correctes faites par rapport à toutes les prédictions faites. Dans cette étude, nous avons 3 classes (BI-RADS 2, BI-RADS 3&4, BI-RADS5), nous avons calculé la précision pour chaque classe à l'aide de la formule suivante :

$$\text{Précision de la classe } i = \frac{\text{Nbr des images correctement attribuées à la classe } i}{\text{Nbr des images attribuées à la classe } i}$$

Pour la précision totale nous avons utilisé la formule :

$$\text{Précision totale} = \frac{\sum_{i=1}^3 \text{Précision de la classe } i}{3}$$

IV.3.1.2 Le rappel ou sensibilité (Recall)

Comme pour la précision nous avons calculé la sensibilité de chaque classe (i) par la formule suivante :

$$\text{Rappel de la classe } i = \frac{\text{Nbr des images correctement attribuées à la classe } i}{\text{Nbr des images appartenant à la classe } i}$$

Et la sensibilité totale par :

$$\text{Rappel de la classe } i = \frac{\sum_{i=1}^3 \text{Rappel de la classe } i}{3}$$

IV.3.1.3 Le F-score

Le F-score noté $F\beta$ -score ou $\beta \in [1, +\infty [$ est un résumé des deux mesures précédentes, Le paramètre β permet d'amplifier l'importance de l'un ou de l'autre des paramètres. Lorsque les deux sont d'équivalente importance, on choisit $\beta = 1$. La formule pour calculer F-Score est :

$$F\beta - \text{score} = (1 + \beta^2) \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\beta^2 \cdot \text{Précision} + \text{Rappel}}$$

Ce score nous donnera le moyen harmonique de précision et de rappel. Mathématiquement, le F1-score est la moyenne pondérée de la précision et du rappel. La meilleure valeur de F1-score serait 1 et la pire serait 0.

IV.3.2 Résultats des expériences

Enfin toute est près pour commencer les expériences, Les paramètres de Haralick des 30 distances sont calculés et bien organisé dans des fichiers « *.csv », le réseau de neurones bien défini et paramétré. Cette tâche consiste à entraîner le réseau de neurones 30 fois, une pour chaque distance, testé chaque modèle obtenu, calculer son précision et son rapport de classification. À la fin nous feront une comparaison entre les 20 résultats obtenus pour avoir la meilleure distance à utiliser dans la construction de la matrice de cooccurrence pour la classification des anomalies mammographiques selon la terminologie BI-RADS. La figure IV.2 présente un exemple d'entraînement du réseau de neurones avec la distance $d=1$.

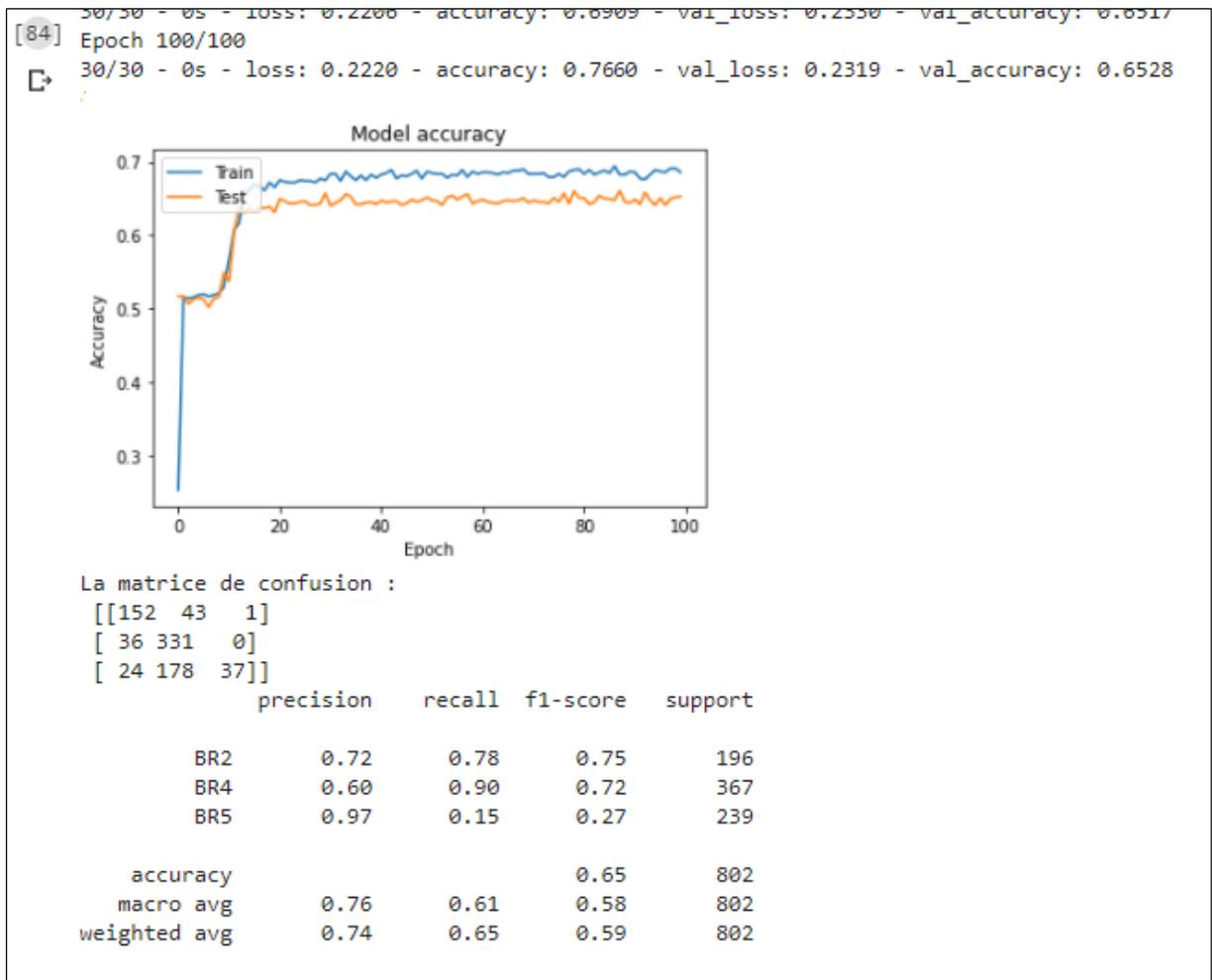


Figure IV.6 Exemple de résultat de l'entraînement du modèle avec les images (distance=1).

Après l'accomplissement de toutes les expériences, les résultats sont regroupés dans le tableau IV.3.

Distances	Précision				Rappel				F1-score				classement
	BR2	BR3-4	BR5	Total	BR2	BR3-4	BR5	Total	BR2	BR3-4	BR5	Total	
1	77.00	60.00	86.00	74.33	71.00	93.00	18.00	60.67	74.00	73.00	29.00	58.67	13
2	74.00	61.00	94.00	76.33	80.00	90.00	19.00	63.00	77.00	73.00	31.00	60.33	11
3	74.00	60.00	85.00	73.00	76.00	91.00	17.00	61.33	75.00	72.00	28.00	58.33	14
4	84.00	61.00	81.00	75.33	75.00	94.00	21.00	63.33	79.00	74.00	34.00	62.33	6
5	85.00	60.00	93.00	79.33	73.00	96.00	17.00	62.00	79.00	74.00	29.00	60.67	10
6	85.00	61.00	88.00	78.00	74.00	96.00	19.00	63.00	79.00	74.00	31.00	61.33	8
7	86.00	61.00	91.00	79.33	76.00	95.00	21.00	64.00	80.00	74.00	33.00	62.33	6
8	85.00	62.00	83.00	76.67	76.00	94.00	24.00	64.67	80.00	75.00	38.00	64.33	1
9	87.00	60.00	88.00	78.33	75.00	96.00	18.00	63.00	81.00	74.00	30.00	61.67	7
10	88.00	60.00	88.00	78.67	74.00	96.00	18.00	62.67	81.00	74.00	30.00	61.67	7
11	87.00	61.00	88.00	78.67	76.00	96.00	21.00	64.33	81.00	75.00	34.00	63.33	4
12	80.00	62.00	88.00	76.67	77.00	93.00	23.00	64.33	79.00	74.00	37.00	63.33	4

13	85.00	62.00	90.00	79.00	76.00	95.00	22.00	64.33	80.00	75.00	36.00	63.67	3
14	84.00	60.00	91.00	78.33	76.00	95.00	17.00	62.67	80.00	73.00	28.00	60.33	11
15	87.00	61.00	85.00	77.67	74.00	96.00	22.00	64.00	80.00	75.00	35.00	63.33	4
16	88.00	60.00	89.00	79.00	74.00	96.00	17.00	62.33	81.00	74.00	28.00	61.00	9
17	86.00	60.00	88.00	78.00	74.00	95.00	18.00	62.33	80.00	74.00	30.00	61.33	8
18	88.00	60.00	89.00	79.00	73.00	96.00	17.00	62.00	80.00	74.00	29.00	61.00	9
19	83.00	60.00	95.00	79.33	78.00	95.00	18.00	63.67	80.00	74.00	30.00	61.33	8
20	87.00	59.00	95.00	80.33	74.00	96.00	16.00	62.00	80.00	73.00	27.00	60.00	12
21	86.00	59.00	93.00	79.33	74.00	96.00	16.00	62.00	80.00	73.00	28.00	60.33	11
22	87.00	60.00	88.00	78.33	74.00	95.00	18.00	62.33	80.00	73.00	29.00	60.67	10
23	87.00	60.00	97.00	81.33	75.00	96.00	16.00	62.33	81.00	74.00	28.00	61.00	9
24	88.00	60.00	85.00	77.67	73.00	95.00	21.00	63.00	80.00	74.00	34.00	62.67	4
25	84.00	53.00	99.00	78.67	45.00	99.00	5.00	49.67	58.00	69.00	10.00	45.67	15
26	89.00	61.00	81.00	77.00	72.00	95.00	24.00	63.67	80.00	75.00	37.00	64.00	2
27	88.00	60.00	89.00	79.00	74.00	96.00	17.00	62.33	81.00	74.00	29.00	61.33	8
28	90.00	60.00	92.00	80.67	72.00	98.00	19.00	63.00	80.00	74.00	31.00	61.67	7
29	90.00	60.00	93.00	81.00	72.00	97.00	18.00	62.33	80.00	74.00	30.00	61.33	8
30	90.00	60.00	93.00	81.00	75.00	97.00	17.00	63.00	82.00	74.00	29.00	61.67	7

Tableau IV.2 Résultats des 30 distances.

En analysant les résultats présentés sur le tableau IV.2 nous pouvons constater que la **distance $d=23$** donne la meilleure précision (81.33%), mais la **distance $d=8$** à la plus grande valeur pour la mesure F1-score. Sachant la mesure F1-score est une combinaison de la précision de la sensibilité, donc elle est plus significative que la précision seule.

Nous avons également remarqué que les résultats obtenus sur ces 30 distances sont presque similaires, ce qui nous conduit à conclure que les petites distances (inférieur à 30) n'ont pas une grande influence sur le résultat de classification des anomalies mammaires.

IV.6 Conclusion

Dans ce dernier chapitre, nous avons présenté l'environnement matériel et logiciel utilisé dans la réalisation de ce travail, que ce soit dans la première partie consacré à la préparation des données d'apprentissage, ou dans la deuxième partie consacré à l'utilisation d'un réseaux de neurones artificiel pour entrainer un modèle avec les 30 distances, et évaluer les performances de classification par rapport aux différentes distance. À la fin, une comparaison entre les différents résultats a été faite pour déterminer la distance à utiliser dans le calcul de la matrice de cooccurrence pour avoir les meilleur performances de classification.

Conclusion Générale

Les systèmes d'aide aux diagnostics (CADx) sont « des applications informatiques dont le but est de fournir aux radiologues en temps et lieux utiles les diagnostics décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des traitements et l'état de santé des patients.

Le travail présenté dans ce mémoire concerne la réalisation d'un system d'aide au diagnostic CADx pour la classification des anomalies mammaires en se basant sur les paramètres d'Haralick. Le but principal est d'analyser le paramètre « d » utilisé pour le calcul de la matrice de cooccurrences GLCM, cette analyse nous permettra de déterminer la meilleure distance « d » à utiliser pour la classification des anomalies mammaires selon la terminologie BI-RADS.

Dans le premier chapitre nous avons présenté un état de l'art sur les méthodes d'analyse des textures pour montrer la position de la méthode à utiliser dans le contexte général du domaine d'étude. Dans le deuxième chapitre nous avons présenté un état de l'art du domaine d'application ainsi que la terminologie utilisée par les radiologues en mammographie pour la classification des anomalies mammaires.

Dans la partie réalisation, nous avons développé une application JAVA pour l'extraction et l'organisation des annotations de la base de données mammographique DDSM. Cette application nous facilitera la sélection des régions d'intérêts (ROIs) et le calcul des quatorze descripteurs de Haralick pour chaque distance « d ».

Un fichier des descripteurs des régions d'intérêts est généré pour chaque distance « d » pour qu'il s'utilise dans la phase d'expérimentations d'apprentissages et des tests du RNA.

Pour l'implémentation du réseau de neurones artificiel nous avons développé une application en utilisant le langage Python, nous avons opté pour Google Colab en utilisant la bibliothèque Keras et TensorFow. Ce réseau de neurones a été testé pour chaque fichier de descripteurs.

Nous avons constaté que la distance $d=23$ donne la meilleure précision (81.33%), mais la distance $d=8$ offre la plus grande valeur pour la mesure F1-score. Sachant la mesure F1-score est une combinaison de la précision de la sensibilité, donc elle est plus significative que la précision seule.

En perspective nous proposons d'explorer les possibilités d'utiliser les techniques d'apprentissage profond pour la classification des anomalies mammaires.

Bibliographie

- [1] B. S. Manjunath et W.-Y. Ma, « Texture features for browsing and retrieval of image data », *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, n° 8, p. 837–842, 1996.
- [2] M. Tuceryan et A. K. Jain, « Texture analysis », in *Handbook of pattern recognition and computer vision*, World Scientific, 1993, p. 235–276.
- [3] F. W. Campbell et J. G. Robson, « Application of Fourier analysis to the visibility of gratings », *The Journal of physiology*, vol. 197, n° 3, p. 551, 1968.
- [4] M. Georgeson, « Spatial Fourier analysis and human vision », *Tutorial Essays in Psychology: Guide to Recent Advances*, p. 39–88, 2014.
- [5] « WHO | Breast cancer », WHO. <http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/> (consulté le 01/10/2020).
- [6] K. I. Laws, « Textured image segmentation », University of Southern California Los Angeles Image Processing INST, 1980.
- [7] W. D. Gillett, « Image Classification Using Laws' Texture Energy Measures », 1987.
- [8] L. Paulhac, « Outils et méthodes d'analyse d'images 3D texturées: application à la segmentation des images échographiques », Thèse de doctorat, 2009.
- [9] D. Gabor, « Theory of communication. Part 1: The analysis of information », *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, n° 26, p. 429–441, 1946.
- [10] S. Mallat, « Image retrieval using wavelet-based salient points [] », *Trans Pattern Anal Mach Intell IEEE*, 1989.
- [11] R. N. Bracewell et R. N. Bracewell, *The Fourier transform and its applications*, vol. 31999. McGraw-Hill New York, 1986.
- [12] M. Unser, « Texture classification and segmentation using wavelet frames », *IEEE Transactions on image processing*, vol. 4, n° 11, p. 1549–1560, 1995.
- [13] I. Fogel et D. Sagi, « Gabor filters as texture discriminator », *Biological cybernetics*, vol. 61, n° 2, p. 103–113, 1989.
- [14] A. Laine et J. Fan, « Texture classification by wavelet packet signatures », *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, n° 11, p. 1186–1191, 1993.
- [15] J. Lim et S. S. Udpa, « Defect extraction in MOI image based on wavelet packet transform and morphological technique », in *AIP Conference Proceedings*, 2001, vol. 557, n° 1, p. 611–618.
- [16] C. Schmid, « Constructing models for content-based image retrieval », in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 2, p. II–II.
- [17] N. E. Huang et Z. Wu, « A review on Hilbert-Huang transform: Method and its applications to geophysical studies », *Reviews of geophysics*, vol. 46, n° 2, 2008.
- [18] G. Rellier, X. Descombes, F. Falzon, et J. Zerubia, « Texture feature analysis using a Gauss-Markov model in hyperspectral image classification », *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, n° 7, p. 1543–1551, 2004.

Bibliographie

- [19] S. Z. Li, *Markov random field modeling in computer vision*. Springer Science & Business Media, 2012.
- [20] X. Descombes, « Champs Markoviens en analyse d'images », Thèse de doctorat, Paris, ENST, 1993.
- [21] G. Winkler, *Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction*, vol. 27. Springer Science & Business Media, 2012.
- [22] J. Besag, « Spatial interaction and the statistical analysis of lattice systems », *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, n° 2, p. 192–225, 1974.
- [23] H. Derin et H. Elliott, « Modeling and segmentation of noisy and textured images using Gibbs random fields », *IEEE Transactions on pattern analysis and machine intelligence*, n° 1, p. 39–55, 1987.
- [24] G. Gimel'farb, *Image textures and Gibbs random fields*. .
- [25] B. B. Mandelbrot, « Fractals », *Form, chance and dimension*, 1977.
- [26] B. B. Mandelbrot, « The Fractal Geometry of Nature/Ed. WH Freeman », *New York*, 1983.
- [27] P. Soille et J.-F. Rivest, « On the validity of fractal dimension measurements in image analysis », *Journal of visual communication and image representation*, vol. 7, n° 3, p. 217–229, 1996.
- [28] S. Peleg, J. Naor, R. Hartly, et D. Avnir, « Multiple resolution texture analysis and classification, IEEE Traits », *Pattern Anal. Mach. Intell., PAM*, vol. 6, n° 4, p. 845–848.
- [29] K. C. Clarke, « Computation of the fractal dimension of topographic surfaces using the triangular prism surface area method », *Computers & Geosciences*, vol. 12, n° 5, p. 713–722, 1986.
- [30] R. F. Voss, « Random fractals: characterization and measurement », in *Scaling phenomena in disordered systems*, Springer, 1991, p. 1–11.
- [31] N. Sarkar et B. B. Chaudhuri, « An efficient differential box-counting approach to compute fractal dimension of image », *IEEE Transactions on systems, man, and cybernetics*, vol. 24, n° 1, p. 115–120, 1994.
- [32] B. Dubuc, S. W. Zucker, C. Tricot, J. F. Quiniou, et D. Wehbi, « Evaluating the fractal dimension of surfaces », *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 425, n° 1868, p. 113–127, 1989.
- [33] J. M. Keller, S. Chen, et R. M. Crownover, « Texture description and segmentation through fractal geometry », *Computer Vision, Graphics, and image processing*, vol. 45, n° 2, p. 150–166, 1989.
- [34] R. M. Haralick, « Statistical and structural approaches to texture », *Proceedings of the IEEE*, vol. 67, n° 5, p. 786–804, 1979.
- [35] R. Vyas, T. Kanumuri, G. Sheoran, et P. Dubey, « Co-occurrence features and neural network classification approach for iris recognition », in *2017 Fourth International Conference on Image Information Processing (ICIIP)*, 2017, p. 1–6.
- [36] R. M. Haralick, « Statistical and structural approaches to texture », *Proceedings of the IEEE*, vol. 67, n° 5, p. 786–804, 1979.

Bibliographie

- [37] S. Njor *et al.*, « Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies », *Journal of medical screening*, vol. 19, n° 1_suppl, p. 33–41, 2012.
- [38] M. Broeders *et al.*, « The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies », *Journal of medical screening*, vol. 19, n° 1_suppl, p. 14–25, 2012.
- [39] I. Thomassin-Naggara, A. Tardivon, et J. Chopier, « Diagnostic et compte rendu standardisé du cancer du sein », *Journal de Radiologie Diagnostique et Interventionnelle*, vol. 95, no 7-8, p. 745–752, 2014.
- [40] I. Cheikhrouhou, « Description et classification des masses mammaires pour le diagnostic du cancer du sein », Thèse de doctorat, 2012.
- [41] E. A. Sickles, « Breast calcifications: mammographic evaluation. », *Radiology*, vol. 160, no 2, p. 289–293, 1986.
- [42] M. Le Gal, G. Chavanne, et D. Pellier, « 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*, vol. 71, n° 1, p. 57–64, 1984.
- [43] I. Bokobsa, L. Levy, M. Suissa, H. Tristant, et M. Benmussa, « Construction et adaptation française de la classification BI-RADS », *27^e Journées de la Société française de sénologie et de pathologie mammaire, Deauville, FRA, 2005-11-16: Dogmes et doutes (revue critique des standards en sénologie)/Dogmas and doubts (critical revue of standards in senology)*, 2005.
- [44] B.-J. Hackelöer, V. Duda, et G. Lauth, *Ultrasound mammography: Methods, results, diagnostic strategies*. Springer Science & Business Media, 2012.
- [45] S. A. Feig *et al.*, « American College of Radiology guidelines for breast cancer screening. », *AJR. American journal of roentgenology*, vol. 171, n° 1, p. 29–33, 1998.
- [46] M. Arnold *et al.*, « Global burden of 5 major types of gastrointestinal cancer », *Gastroenterology*, 2020.
- [47] A. Mahnane et M. H. CHERIF, « Epidémiologie du cancer du sein en Algérie », *Actualités dans la prise en charge multidisciplinaires des cancers du sein en*, 2012.
- [48] D. Droixhe, « Raconter le cancer du sein. Le traitement par la ciguë de deux religieuses e Pau en 1759-1761 », *Revue de Pau et du Bearn*, vol. 42, 2015.
- [49] E. R. Fisher, R. M. Gregorio, C. Redmond, et B. Fisher, « Tubulolobular invasive breast cancer: a variant of lobular invasive cancer », *Human pathology*, vol. 8, n° 6, p. 679–683, 1977.
- [50] Topsante.com, « Quand, comment et où passer une mammographie ? - Top Santé », mars 02, 2017. <https://www.topsante.com/themes/mammographie> (consulté le 01/10/2020).
- [51] A. Kaehler et G. Bradski, *Learning OpenCV 3: computer vision in C++ with the OpenCV library*. O'Reilly Media, Inc., 2016.
- [52] « USF Digital Mammography Home Page ». <http://www.eng.usf.edu/cvprg/Mammography/Database.html> (consulté le 01/10/2020).
- [53] « Apache Derby ». <http://db.apache.org/derby/> (consulté le 01/10/2020).

Bibliographie

- [54] « Introduction to the Java Persistence API - The Java EE 6 Tutorial ». <https://docs.oracle.com/javaee/6/tutorial/doc/bnbpz.html> (consulté le 01/10/2020).
- [55] « TensorFlow Core », *TensorFlow*. <https://www.tensorflow.org/tutorials?hl=fr> (consulté le 01/10/2020).
- [56] « Keras: the Python deep learning API ». <https://keras.io/> (consulté le 01/10/2020).
- [57] « NumPy ». <https://numpy.org/> (consulté le 01/10/2020).
- [58] « scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation ». <https://scikit-learn.org/stable/> (consulté le 01/10/2020).
- [59] « Cancer du sein - Ooreka », *Ooreka.fr*. [//cancer-du-sein.ooreka.fr](https://cancer-du-sein.ooreka.fr) (consulté le 01/10/2020).
- [60] « Cours ». http://campus.cerimes.fr/maieutique/UE-gynecologie/examen_praclinique/site/html/10.html (consulté le 01/10/2020).

Annexes

Annexe 01 : *Le script python « convert.py »*

```
1 import os
2 path_to_ddsm = "./DDSM"
3 for root, subFolders, file_names in os.walk(path_to_ddsm):
4     for file_name in file_names:
5         if ".LJPEG" in file_name:
6             ljpeg_path = os.path.join(root, file_name)
7             out_path = os.path.join(root, file_name)
8             out_path = out_path.split('.LJPEG')[0] + ".jpg"
9             cmd = './ljpeg.py "{0}" "{1}" '.format(ljpeg_path, out_path)
10            os.system(cmd)
11
12 print('done')
13
```

Annexes

Annexe 02 : Le script Python « ljpeg.py »

```
#!/usr/bin/env python
import os
import sys
import re
import subprocess
import numpy
import argparse

BIN = os.path.join(os.path.dirname(__file__), "jpegdir", "jpeg")
if not os.path.exists(BIN):
    print "logiciel jpeg introuvable "
    sys.exit(0)

# exemple de output retourne par logiciel stanford ljpeg
#> GW:1979 GH:4349 R:0
#>> C:1 N:nom_du_fichier_décompressé.ljpeg.1 W:1979 H:4349 hf:1 vf:1

PATTERN = re.compile('\sC:(\d+)\s+N:(\S+)\s+W:(\d+)\s+H:(\d+)\s')

def lire_img (path):
    cmd = '%s -d -s %s' % (BIN, path)
    l = subprocess.check_output(cmd, shell=True)
    #print l
    m = re.search(PATTERN, l)
    C = int(m.group(1)) # C=1 represente le canal
    F = m.group(2) # le nom du fichier
    W = int(m.group(3)) # width
    H = int(m.group(4)) # Hight
    im = numpy.fromfile(F, dtype='uint16').reshape(H, W)
    L = im >> 8
    H = im & 0xFF
    im = (H << 8) | L
    os.remove(F)
    return im

if __name__ == '__main__':
    import glob
    import cv2
    parser = argparse.ArgumentParser()
    parser.add_argument("ljpeg", nargs=1)
    parser.add_argument("output", nargs=1)
    args = parser.parse_args()
    path = args.ljpeg[0]
    tiff = args.output[0]
    root = os.path.dirname(path)
    # lire ICS
    ics = glob.glob(root + '/*.ics')[0]
    name = path.split('.')[ -2]
    W = None
    H = None
    # trouver la dimension de l'image dans ics
    for l in open(ics, 'r'):
        l = l.strip().split(' ')
        if len(l) < 7:
            continue
        if l[0] == name:
            W = int(l[4])
            H = int(l[2])
            bps = int(l[6])
    image = lire_img(path)
    if W != image.shape[1]:
        image = image.reshape((H, W))
    raw = image
    image = cv2.normalize (image, None, 0, 255, cv2.NORM_MINMAX) #normalisation
    image = numpy.uint8 (image)
    cv2.imwrite(tiff, image)
```

Annexe 03 : Un script Python pour calculer les paramètres de Haralick.

```
1 # function to extract haralick textures from an image
2 import cv2
3 import numpy as np
4 import os
5 import glob
6 import mahotas as mt
7 import csv
8 def extract_features(image,d):
9     textures = mt.features.haralick(image,distance =d)
10    ht_mean = textures.mean(axis=0) # take the mean of it and return it
11    strht=[]
12    for x in ht_mean:
13        strht.append(x)
14    strht.append(cur_label)
15    c.writerow( strht )
16    return ht_mean
17
18 # Load the training dataset
19 train_path = "/content/drive/My Drive/DDSM-BR-28092020"
20 train_names = os.listdir(train_path)
21 print (train_names)
22 i = 1
23 for i in range(1,21) : # Loop over the matrix distance
24    csv_file="/content/drive/My Drive/haralic_D"+str(i)+".csv"
25    c = csv.writer(open(csv_file, "w" ))
26    print ("[Status] preparaton extraction haralick textures..")
27    for train_name in train_names: # Loop over the training classes
28        cur_path = train_path + "/" + train_name
29        cur_label = train_name
30        for file in glob.glob(cur_path + "/*.JPG"): # Loop over the training dataset
31            print ("Processing Image - {} in {}".format(file, cur_label))
32            image = cv2.imread(file) # read the training image
33            gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY) # convert the image to gray:
34            # extract haralick texture from the image
35            features = extract_features(gray,i)
```

Activ
Accé

Annexes

Annexe 04 : Le script Python pour modéliser le réseau de neurones artificiel.

```
1 # neural network with keras
2 from sklearn.model_selection import train_test_split
3 #from sklearn.metrics import confusion_matrix
4 from sklearn.metrics import classification_report, confusion_matrix
5 import numpy as np
6 import tensorflow.compat.v2 as tf
7 #tf.disable_v2_behavior()
8 from keras.models import Sequential
9 from keras.layers import Dense
10 from keras.layers import Dropout
11 #import csv
12 from pandas import read_csv
13 import keras
14 url = "/content/drive/My Drive/MONFICHER11sans35.csv"
15 names = ['f1', 'f2', 'f3', 'f4', 'f5', 'f6', 'f7', 'f8', 'f9', 'f10', 'f11', 'f12', 'f13']
16 dataset = read_csv(url, sep=';')
17 array = dataset.values
18 X = array[:,0:13]
19 y = array[:,13]
20 one_hot_y = keras.utils.to_categorical(y,num_classes =3)
21 X_train, X_validation, Y_train, Y_validation = train_test_split(X, one_hot_y, test_size=0.2)
22 model = Sequential() # define the keras model
23 model.add(Dense(26, input_dim=13,activation='relu', use_bias=False))
24 model.add(Dense(26,activation='relu', use_bias=False))
25 model.add(Dense(26,activation='relu', use_bias=False))
26 model.add(Dense(3,activation='sigmoid'))
27 model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
28 history = model.fit(X_train , Y_train , validation_split=0.2, epochs=130, batch_size=32)
29 Yp=model.predict(X_validation,Y_validation.any() )
30 loss,accuracy = model.evaluate(X_validation, Y_validation,verbose=0)
31 print('Accuracy: %.2f' % (accuracy*100))
32 #-----
33 import matplotlib.pyplot as plt
34 # Plot training & validation accuracy values
35 plt.plot(history.history['accuracy'])
36 plt.plot(history.history['val_accuracy'])
37 plt.title('Model accuracy')
38 plt.ylabel('Accuracy')
39 plt.xlabel('Epoch')
40 plt.legend(['Train', 'Test'], loc='upper left')
41 plt.show()
42
43 # Plot training & validation loss values
44 plt.plot(history.history['loss'])
45 plt.plot(history.history['val_loss'])
46 plt.title('Model loss')
47 plt.ylabel('Loss')
48 plt.xlabel('Epoch')
49 plt.legend(['Train', 'Test'], loc='upper left')
50 plt.show()
51
52 cm = confusion_matrix(Y_validation.argmax(axis=1),Yp.argmax(axis=1) )
53 print(cm)
54 target_names = ['BR2', 'BR4', 'BR5']
55 print(classification_report(Y_validation.argmax( axis=1), Yp.argmax(axis=1),target_names=target_names))
```