République algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

Université Hamma Lakhdar. El-Oued
Faculté des Sciences Exactes

Département D'informatique

Mémoire de fin d'études

MASTER ACADEMIQUE

Domaine : Mathématiques et Informatique Filière : Informatique

Spécialité : Systèmes Distribués et Intelligence Artificiel (SDIA)

Présenté par :

HADFI Chiraz TOUANSA Insaf

Thème

Analyse de la texture des images mammaires par une fusion des lois de Zipf et des filtres Gabor dans un processus de classification des tumeurs mammaires

discuté devant le jury composé de:

Dr	**Hamoud Meriem**	**Superviseur**	**Univ. El Oued**
Dr	**Ben Ali Abdelkamel**	**Président**	**Univ. El Oued**
M.	**retimea farida**	**Examinateur**	**Univ. biskra**

Année universitaire : 2019/2020

Remerciements

Nous tenons à remercier du fond du cœur, et par-dessus tout, Dieu Tout-Puissant de nous avoir donné la force, la volonté et le courage de faire ce travail.

Nous tenons à remercier chaleureusement et successivement tous ceux qui ont contribué de près et de loin à la réalisation de ce projet d'étude final. Nous remercions la Dr Maryam Hammoud pour sa confiance et son encadrement efficace et lui exprimons notre profonde gratitude pour le partage de ses connaissances et de sa gentillesse, ainsi que de ses méthodes de travail, notamment pour sa rigueur scientifique.

Nous remercions les membres du jury d'avoir accepté de juger nos travaux.

Nous exprimons également notre gratitude à tous ceux qui ont coopéré à

notre formation, en particulier les enseignants du Département des technologies de l'information de l'Université Shahidat Hama Al-Akhdar Al-Wadi pour nos collègues du semestre 2019-2020. Nous remercions également tous ceux qui ont participé directement ou indirectement au développement de ce travail.

Dédicaces

Je dédie ce travail

À mes chers parents *pour leurs sacrifices, leur amour, leur soutien et leurs prières tout au long de mes études. Ils peuvent trouver ici un témoignage de ma profonde gratitude,*

À mes chères sœurs *Rahma, djoumana , Amina et Amani pour leurs encouragements constants et leur soutien moral,*

À tous mes amis *pour les aider et les soutenir dans les moments difficiles.*

Merci à tous d'être toujours à mes côtés.

Que Dieu vous accorde la santé, le bonheur, le courage et, par-dessus tout, le succès.

Insaf

Dédicaces

Je dédie ce travail

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

A mes chères sœurs …… pour leurs encouragements permanents, et leur soutien moral,

A mes chers frères, ……., pour leur appui et leur encouragement,

A toute ma famille pour leur soutien tout au long de mon parcours universitaire,

À mon époux qui ne manque pas à me donne de courages sans arrêt.

À mes chères amies, à tous ceux qui m'aiment et à tous ceux que j'aime.

Merci d'être toujours là pour moi.

chiraz

Résumé

Vu la difficulté de modéliser les structures complexes des images via de simples relations linéaires; la non-linéarité pourrait être une alternative. Dans ce sens, notre problématique de recherche est fondée sur un ultime but de contribuer une technique robuste d'analyse et de caractérisation de la texture au sein des images mammaires durant un processus de classification de lésions. En effet, l'enjeu majeur de notre approche réside dans l'analyse de la texture, à la fois dans le domaine spatial (les lois puissance : Zipf et de Zipf inverse) et dans le domaine fréquentiel (les filtres de Gabor). Les approches statistiques utilisent les propriétés qui régissent la distribution et les relations des niveaux de gris dans l'image. D'autre part, les méthodes basées transformation génèrent les caractéristiques de texture à différentes fréquences et orientations. Par la suite, nous ferons apprendre les algorithmes d'apprentissage automatique sur des régions d'intérêt segmentées et étiquetées en deux classes : tumeurs malignes ou bénignes et tissu normal ou tumeur ; pour trouver la meilleure hypothèse qui mappe les données d'entrée à la sortie souhaitée. Eventuellement, nous implémenteront cette approche via une technique d'aide au diagnostic médical du cancer du sein assisté par ordinateur (CAD), qui servira comme un deuxième avis aux radiologues, où des résultats satisfaisants ont été obtenus.

Mots clés : Analyse d'images, Vision par ordinateur, Mammographie, Tumeur, CAD, Loi de Zipf, Loi de Zipf inverse, Filtres de Gabor, Apprentissage automatique, Classification.

Abstract

Given the difficulty of modeling the complex structures of images through simple linear relations; non-linearity could be an alternative. In this sense, our research problem is based on the ultimate goal of contributing a robust technique of texture analysis and characterization within breast images during a lesion classification process. Indeed, the major challenge of our approach lies in the analysis of the texture, both in the spatial domain (the power laws: Zipf and reverse Zipf) and the frequency domain (the Gabor filters). Statistical approaches use the properties that govern the distribution and relationships of gray levels in the image. On the other hand, transformation-based methods generate texture characteristics at different frequencies and orientations.

Next, we train machine learning algorithms on regions of interest segmented and labeled into two classes: malignant or benign tumors and normal parenchyma tissue or tumor; to find the best hypothesis that maps the input data to the desired output. Eventually, we will implement this approach via a technique to aid in the medical diagnosis of breast cancer (CAD) which should provide a second opinion to assist the radiologist's decision, where satisfactory results have been obtained.

Keywords: Image analysis, Computer vision, Mammography, Tumor, CAD, Zipf's law, inverse Zipf's law, Gabor filters, Machine learning, Classification.

الملخص

بالنظر إلى صعوبة نمذجة الهياكل المعقدة للصور عبر العلاقات الخطية البسيطة؛ يمكن أن تكون العلاقات اللاخطية بديلاً. بهذا المعنى، تستند مشكلة البحث لدينا على الهدف النهائي للمساهمة بتقنية قوية لتحليل وتوصيف النسيج داخل صور الثدي أثناء عملية تصنيف الورم. في الواقع، تكمن المصلحة الرئيسية لنهجنا في تحليل النسيج، سواء في المجال المكاني (قوانين القوة: زيب اف وزيب اف العكسي) وفي مجال التردد (مرشحات غابور). تستخدم المناهج الإحصائية الخصائص التي تحكم توزيع وعلاقات مستويات الرمادي في الصورة. من ناحية أخرى، تولد الأساليب القائمة على التحويل خصائص نسيجية في ترددات واتجاهات مختلفة.

بعد ذلك، سنقوم بتدريس خوارزميات التعلم الآلي في مناطق الاهتمام مقسمة ومُصنفة إلى فئتين: الأورام الخبيثة والحميدة للعثور على أفضل الفرضية التي ترسم بيانات الإدخال إلى الناتج المطلوب. في النهاية، سنقوم بتنفيذ هذا النهج من خلال تقنية للمساعدة في التشخيص الطبي لسرطان الثدي بمساعدة الكمبيوتر، والتي ستكون بمثابة رأي ثانٍ لأخصائي الأشعة، حيث تم الحصول على نتائج مرضية.

كلمات البحث: تحليل الصور، رؤية الكمبيوتر، التصوير الشعاعي للثدي، الورم، قانون زيب اف، قانون زيب اف العكسي، مرشحات غابور، التعلم الآلي، التصنيف.

Table des matières

Chapitre III : L'apprentissage automatique en action : état de l'art sur les approches assistées par ordinateur de pronostic du cancer du sein

Chapitre VI : Problématique, approche proposée et évaluation des résultats obtenus

Table des figures

Chapitre II : Analyse des images par les lois de Zipf et de Zipf inverse

Chapitre III : L'apprentissage automatique en action : état de l'art sur les approches assistées par ordinateur de pronostic du cancer du sein

Chapitre VI : Problématique, approche proposée et évaluation des résultats obtenus

Liste des tableaux

Chapitre II : Analyse des images par les lois de Zipf et de Zipf inverse

Chapitre VI : Problématique, approche proposée et évaluation des résultats obtenus

LISTE DES ACRONYMES

3D	trois dimensions
BPNN	Back Propagation Neural Network
CAD	Computer Aided Design
CADe	Computer Aided Detection
CADx	Computer Aided Diagnosis
CC	Cranial-Caudal
DDSM	Digital Database for Screening Mammography
DSIFT	Dense Scale Invariant Feature Transform
FN	False Negative
FP	False Positive
GLCM	Gray-Level Co-occurrence Matrix
GLRLM	Grey-Level Run Length Matrix
HOG	histogram of oriented gradients
IA	l'Intelligence Artificielle
KNN	K Nearest Neighbors
LBP	Local Binary Patterns
LCP	longest common prefix
LDA	linear discriminant analysis
LDA	Latent Dirichlet Allocation
MATLAB	MATrix LABoratory
MIAS	Mammographic Image Analysis Society
ML	Machine Learning
MLO	Medum Lateral Lblique
MLP	MultiLayer Perceptron
PCA	Principal Component Analysis

QDA	Quadratic Discriminant Analysis
ROIs	Region Of Interest
SGLD	Stochastic gradient Langevin dynamics
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WGLCM	Wavelet Gray Level Co-occurrence Matrix

Introduction générale

Le cancer du sein s'est avéré le cancer le plus répandu dans la population femme du monde entier, en effet, jusqu-à ce jour, uniquement la détection précoce de cette maladie dangereuse peut booster la capacité du traitement ainsi que le taux de survie des patientes atteintes, ceci via la mammographie de dépistage.

La tendance croissante à l'application des techniques d'apprentissage automatique pour la prédiction du pronostic du cancer est principalement reliée au fait que de grandes bases de données composées de données complexes provenant de divers domaines pourraient être efficacement explorées en peu de temps. En effet, l'apprentissage automatique est un domaine de recherche très actif en intelligence artificielle où l'ultime but est de doter les machines de l'anthropomorphe capacité d'apprendre, de raisonner ainsi que de prendre des décisions en référence à une classe d'algorithmes informatiques qui construisent des modèles pour la classification.

Par la suite, le résultat de la phase d'apprentissage sera un ensemble de paramètres optimisés sous forme d'un modèle effectuant des tâches de prédiction où l'ordinateur apprend par expérience, et en utilisant cette expérience améliore ses performances.

L'apprentissage automatique fut la base des systèmes de détection ou de diagnostic assistés par ordinateur (CAD) développés pour aider les radiologues à améliorer la précision du diagnostic, en leur donnant un deuxième avis pour l'établissement du diagnostic final. Dans ce sens, la problématique principale de ce travail concerne la contribution de descripteurs discriminants qui quantifient les informations texturales des tumeurs mammaires. En effet, bien que chaque type de tissu ait ses propres caractéristiques, il demeure difficile de distinguer entre le tissu cancéreux et le tissu bénin. A cet effet, nous contribuons les lois puissance: Zipf et Zipf inverse pour l'analyse de la texture des images mammaires. Plus précisément, nous élaborons une fusion des lois de Zipf et de Zipf inverse avec l'approche d'analyse de texture des filtres de Gabor sous l'ultime but d'explorer l'apport complémentaire que peut révéler ces deux approches de caractérisation de la texture à plusieurs résolutions et procédant de manière analogue que le système visuel humain, ceci, à travaers la contribution d'une approche de diagnostic du cancer du sein assitée par ordinateur. Donc, les lois de puissance seront exploitées pour décrire les dépendances spatiales entre les pixels et nous effectuons en parallèle l'analyse fréquentielle de l'image par les filtres de Gabor.

Ce mémoire est structuré comme suit :

Chapitre I : Analyse d'image et vision par ordinateur

Dans le chapitre I, les notions et concepts de base de l'analyse d'image ainsi que de la vision par ordinateur sont présentées, tout en évoquant l'analyse de la texture des images.

Chapitre II : Analyse des images par les lois de Zipf et de Zipf inverse

Dans le chapitre II, nous nous focalisons sur les lois puissance: Zipf et Zipf inverse en analyse d'image.

Chapitre III : L'apprentissage automatique en action : état de l'art sur les approches assistées par ordinateur de pronostic du cancer du sein

Nous introduirons dans le troisième chapitre, le cancer du sein ainsi que les aspects fondamentaux de l'apprentissage automatique en évoquant certaines des avancées les plus récentes dans le domaine de l'apprentissage automatique et de l'aide au diagnostic médical du cancer du sein assisté par ordinateur.

Chapitre IV : Problématique, approche proposée et évaluation des résultats obtenus

Au long du quatrième chapitre, nous présentons la problématique tirée ainsi que l'approche proposée. Evidement ; une validation des résultats obtenus est prévue.

Ce travail s'achèvera par une conclusion générale résumant les pivots de la contribution proposée, suivie par les perspectives soulignées.

Chapitre I

Analyse d'image et vision par ordinateur

1. Introduction

L'analyse est une étape très importante en imagerie. En effet, les principales informations pour l'interprétation d'un message visuel pour un observateur humain demeurent les contours ou bien les textures.

L'analyse de l'image désigne l'extraction d'un nombre de caractéristiques et à les exprimer sous forme paramétrique. Cette étape d'extraction des paramètres précède généralement une phase de décision pour répondre à des questions pertinentes telles que: est-ce un matériau normal ou défectueux? Ou plus précisément dans notre domaine de recherche : est-ce un tissu biologique sain ou bien pathologique?

La texture est un attribut visuel incontournable dans la vision par ordinateur. En effet, la problématique majeure réside dans le fait qu'il est facile pour le système visuel humain de distinguer les structures de texture mais que ceci présente un défi quant aux méthodes automatiques.

Nous présentons dans le premier chapitre les concepts de base de l'analyse d'image et de la vision par ordinateur. De plus, nous nous focalisons sur la définition de l'analyse numérisée de la texture au sein des images pour la prise de décision.

2. L'analyse d'image

Une image est un signal numérique 2D pouvant être interprétée comme une matrice 2D de valeurs lumineuses ou de couleur. A cet effet, nous distinguons diverses interprétations des images.

En effet, les images peuvent être vues comme un ensemble de pixels, caractérisés par leur intensité lumineuse qui a tendance à varier, représentant une scène concrète ou abstraite. Elles peuvent être modélisées comme une fonction f dont les valeurs f(x,y) désignent les coordonnées spatiales.

Une image est aussi un ensemble de textures ou motifs agencés d'une manière spécifique permettant la description d'une scène. Ces différentes visions sont autant de bases de représentation ; ceci en valorisant différents types d'informations qui aboutissent à différents types d'analyse des images numériques [1] [2].

L'analyse d'image de bas niveau et la vision par ordinateur de haut niveau manipulent diverses données. En effet, les données de bas niveau représentent des images d'origine sous forme de matrices composées de valeurs d'intensité lumineuse. Si nous évoquons l'utilité de l'analyse d'image de bas niveau, nous pourrons l'illustrer par le domaine de la radiographie où l'analyse de bas niveau agit pour aboutir à une visibilité optimale des organes au sein de l'image et permettre ainsi de lever des indécisions [3][4].

D'autre part, les données de haut niveau sont à leur tour originaires des images, cependant, uniquement les données jugées pertinentes et se rapportant à des objectifs de haut niveau seront extraites, évidemment, une réduction considérable de la quantité de données sera observée. L'appel à l'intelligence artificielle est en ordre pour manipuler des connaissances sur les données extraites dans le but d'interpréter ce que représentent les images traitées.

Nous allons évoquer les niveaux d'analyse d'une image comme suit [5][6] :

2.1. Analyse de bas niveau d'image

Les techniques de bas niveau de l'analyse d'image désignent la base du traitement numérique des images en faisant appel à des connaissances minimes concernant le contenu des images.

L'échantillonnage consiste en le procédé de discrétisation spatiale d'une image visant à attribuer à chaque région rectangulaire R(x,y) relative à une image continue, une valeur I(x,y) pour la résolution spatiale. D'autre part, la quantification consiste en la sélection du bon nombre de bits pour la codification des images numériques. Ceci, en donnant une valeur numérique spécifique

à chaque échantillon prélevé sur le signal tout en considérant la contrainte de la réduction du nombre de bits désigné durant le processus du codage.

Une fois la numérisation est effectuée, les images pourront être visualisées sur un moniteur en vue de les traiter ou les stocker ou bien les transmettre à travers un réseau informatique.

Nous distinguons également les traitements effectués sur les images au plus bas niveau d'abstraction en éliminant les indésirables distorsions ainsi qu'en améliorant certaines caractéristiques pertinentes de l'image où nous pouvons citer l'élimination du bruit acquis dans l'image, l'amélioration du contraste des objets pertinents lors du processus d'interprétation de l'image. Ce processus intervient essentiellement dans le cas des images médicales à faible contraste et à un taux élevé de bruit. Un exemple d'un traitement que nous pourrions effectuer sur une image consiste en la restauration d'un pixel déformé durant le processus de numérisation ; sachant que les pixels voisins relatifs à un objet situé dans une image ont essentiellement la même valeur d'intensité lumineuse, de manière que si un pixel déformé peut être capté â partir de l'image, il pourra être restauré sous la forme d'une valeur moyenne des pixels voisins.

Citons aussi le processus de segmentation d'image appartenant à ce bas niveau d'analyse d'image dont le but est de séparer les objets de l'arrière-plan de l'image. Evidemment, seules les caractéristiques qui interviendront lors de l'analyse ultérieure de haut niveau sont extraites. La description d'objet et l'extraction de caractéristiques discriminantes dans une image totalement segmentée est considérée également comme une analyse de bas niveau d'image.

2.2. Analyse de haut niveau d'image

Le traitement de haut niveau ou la reconnaissance d'objet consiste à attribuer une classe à un objet en s'appuyant sur des connaissances relatives du contenu de l'image. Des méthodes d'intelligence artificielle sont incorporées dans le domaine de l'analyse et l'interprétation de l'image. En effet, l'analyse d'image de haut niveau s'appuie sur l'imitation de la cognition humaine.

3. La vision par ordinateur

La vision par ordinateur ou la vision artificielle consiste en une initiative d'automatisation de diverses taches de la vision humaine sur un ordinateur. Considérée comme une discipline scientifique, la vision par ordinateur est bronchée à la théorie des systèmes artificiels visant à

extraire des informations à partir d'images sous diverses formes : des séquences vidéo, des données multi-dimensionnelles depuis un scanner médical.

Si nous observons la conception moderne de l'Intelligence Artificielle (IA), nous affirmerons que la machine perçoit son environnement tout en s'y adaptant. En effet, le traitement d'image ainsi que la vision par ordinateur se basent sur des connaissances et des techniques d'intelligence artificielle dans la gestion et la prise de décision [7].

Ces théories sont appliquées dans la conception des systèmes de vision artificielle, citons quelques exemples [8] :

- Contrôle de processus (un robot industriel).
- Détection de mouvement (vidéo-surveillance).
- Organisation de l'information (indexation des bases de données d'images).
- Analyse de l'imagerie médicale constituant l'outil principal de la vision par ordinateur (systèmes d'aide au diagnostic médical assistés par ordinateur).

4. Analyse de la texture d'image

L'analyse de la texture est une problématique de base dans le traitement d'image et la vision par ordinateur. En effet, elle demeure un problème clé dans plusieurs domaines d'application, tels que la reconnaissance d'objets, la télédétection, la recherche d'image par le contenu et spécialement l'imagerie médicale.

4.1. Définition de la texture

A ce jour, il n'existe aucune définition précise et universelle du concept de la texture. En effet, les textures naturelles sont très irrégulières et restent sans définition précise, malgré leur omniprésence dans les images (images médicales, aériennes, de textiles).

Nous évoquons les définitions suivantes :

La définition donnée par le dictionnaire, affirme qu'une texture est la reproduction spatiale d'un motif de base dans plusieurs directions.

Selon [9], la texture est une propriété de la surface ou de la structure de l'objet. Les caractéristiques de texture extraites sont bénéfiques pour reconnaître et distinguer diverses structures.

Une texture est une répétition d'éléments avec une certaine fréquence et caractérisée par différentes statistiques (moyenne, variance, histogramme,…) [10].

Dans [11], une texture est définie telle qu'une région d'une image présentant une organisation spatiale homogène des niveaux de luminance.

Nous tirons profit de ces définitions pour affirmer au long de ce mémoire, que la texture est un descripteur incontournable dans le domaine de traitement d'images. En effet, elle est déterminante pour la reconnaissance des objets en ayant un vaste champ d'applications dans les processus de segmentation et de classification d'images.

4.2. Typologie de la texture

Nous distinguons deux types de textures [12] :

- Les textures déterministes (ou bien périodiques, ou aussi macrotextures), désignant une répartition spatiale régulière du motif.
- Les textures probabilistes (ou bien stochastiques, aléatoires, ou bien microtextures), relatives à une répartition spatiale irrégulière et aléatoire et ayant divers motifs impossible à isoler ou séparer.

4.3. Caractérisation de la texture d'image

Les méthodes d'analyse et de caractérisation de la texture peuvent être divisées en quatre catégories : [11][13]

- **Méthodes statistiques**

Les méthodes statistiques se fondent la plupart du temps sur la distribution spatiale des niveaux de gris des pixels et sur la description statistique de leur arrangement. Elles consistent à extraire à l'aide des outils statistiques, des paramètres texturaux. Dans ce sens, les matrices de cooccurrence d'Haralick [14] sont les plus fréquemment utilisées.

- **Méthodes géométriques :**

Ce type de méthodes tient compte de l'information structurelle et contextuelle de l'image. Ces méthodes sont particulièrement bien adaptées aux textures macroscopiques (structurelles). En effet, la description de la texture est faite par une extraction explicite de primitives (primitives : ensemble connexe de pixels qui partagent des propriétés similaires) et des règles de placement de ces primitives, par le biais d'attributs appelés attributs géométriques. Cependant, comme ces attributs géométriques sont sensibles à la régularité des motifs texturés présents dans l'image, ils ne peuvent pas caractériser des textures irrégulières comme celles présentes dans la majorité des images naturelles.

- **Méthodes basées sur le modèle :**

Ces méthodes considèrent la texture comme la réalisation d'un processus stochastique stationnaire. Elles se fondent sur la recherche d'un modèle pour décrire ou générer une texture. Les méthodes à base de modèles probabilistes sont largement utilisées, les plus connues sont le modèle autorégressif, le modèle Markovien et le modèle, les champs aléatoires Markoviens gaussiens, les fractales.

- **Les méthodes fréquentielles**

Ces méthodes exploitent le fait que le système visuel humain réalise des analyses fréquentielles de l'image. Celles les plus fréquemment utilisées sont les filtres de Gabor, la transformée de Fourier, les transformées en ondelettes. Les méthodes d'analyse par banc de filtre appliquent une série de filtres à l'image, chacun d'eux permet de mettre une texture de fréquence et d'orientation bien spécifique.

4.4. Problématiques d'analyse de texture

Nous avons différentes problématiques d'analyse de la texture comprenant: la classification de texture, la segmentation de texture, la détermination d'une forme par la texture ainsi que la synthèse d'une texture [15].

- **Classification de texture**

  Dans la classification de texture, l'ultime but consiste à assigner un ensemble de textures variées et inconnue à l'une des classes prédéfinies préalablement. L'affectation ou l'attribution s'effectue sur la base de règles dérivées automatiquement suite à l'analyse d'un ensemble d'apprentissage composé d'échantillons de texture avec des classes connues [16].

- **Segmentation de texture**

  Dans la segmentation basée analyse de texture, le principe consiste à diviser une image à des régions cohérentes. Notons que dans la segmentation supervisée de texture, nous fournissons au système des modèles de textures dans le but d'être rencontrés et reconnus dans les images à segmenter. En revanche, la segmentation non-supervisée de texture divise une image en régions de textures similaires sans avoir d'information a priori sur les textures contenues dans l'image.

- **Détermination d'une forme par la texture**

  L'enjeu de la détermination d'une forme basée sur la texture consiste à deviner la forme de trois dimensions (3D) d'un objet depuis de son image. En effet, la texture est un descripteur pertinent dans la perception de la forme 3D.

- **Synthèse de la texture**

  La synthèse de texture consiste à synthétiser plusieurs échantillons de textures similaires de façon perceptuelle.

5. Conclusion

Dans l'analyse d'image et la vision par ordinateur, la texture présente un descripteur très important pour la description du contenu de l'image. Dans ce sens, de nombreuses approches d'analyse de la texture ont été développées et validé à travers les travaux de la littérature.

Durant un processus de classification, une bonne précision obtenue est directement liée aux caractéristiques des descripteurs et à leur capacité à discriminer les images dans les différentes classes. A cet effet, les chercheurs sont toujours à la recherche de descripteurs faciles à calculer tout en ayant le pouvoir de discriminer les classes, pour contribuer considérablement à la performance de classification.

Nous suggérons une caractérisation judicieuse de la texture des images mammaires pour un processus ultérieur de classification. Pour ceci, nous allons fusionner l'approche des filtre de Gabor avec celle des lois puissance Zipf et Zipf inverse pour la caractérisation de la texture des zones d'intérêt extraites à partir d'images mammaires dans un processus d'aide au diagnostic médical du cancer du sein assisté par ordinateur.

Durant le chapitre suivant ; nous introduirons les lois puissance de type: Zipf et Zipf inverse ; spécialement leur application en analyse d'image et vision par ordinateur.

Chapitre II

Analyse des images par les lois de Zipf et de Zipf inverse

1. Introduction

Tout au long de ce chapitre, nous introduirons les lois de puissance de type Zipf et Zipf inverse pour la modélisation de phénomènes extrêmement différents. Evidemment, nous nous intéressons à l'application de ces approches au domaine de l'analyse d'image et la vision par ordinateur où il a été prouvé, d'une part, qu'une relation existe entre les fréquences d'apparition des différentes occurrences d'un phénomène et le rang respectifs de ces occurrences dans une suite ordonnée. D'autre part, il a été observé une relation existante aussi entre les fréquences d'apparitions des attributs et le nombre d'attributs ayant la même fréquence d'apparition.

2. Qu'est-ce qu'une loi puissance ?

Les lois puissance, citons la loi de Pareto ou les lois de Zipf ont été impliquées dans la modélisation de divers phénomènes réels distincts. Effectivement, des distributions en lois puissance ont été observées dans des domaines scientifiques comme : la physique, la biologie, la psychologie, la sociologie, l'économie, la linguistique) [5].

Une loi puissance peut être définie comme une fonction qui lie deux quantités et prenant la forme $y = ax^{-b}$, où a et b sont des constantes. Sa représentation graphique dans un repère (x,y) est la suivante :

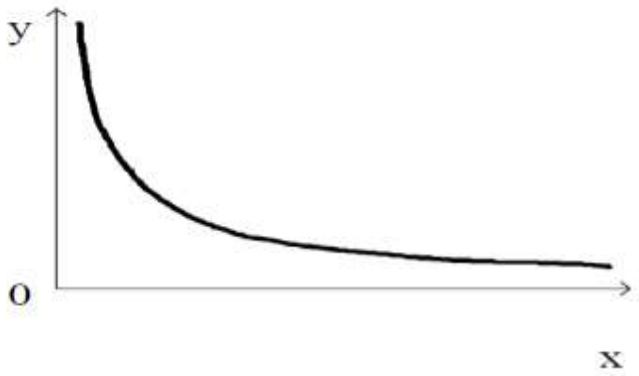

Figure 1:Représentation d'une loi puissance dans un repère linéaire.

Si nous introduisons le logarithme à chaque membre de l'égalité $y = ax^{-b}$, ceci conduira à l'expression : log(y) = log(a) − b log(x).

La représentation graphique d'une loi puissance dans un repère bi-logarithmique est la suivante :

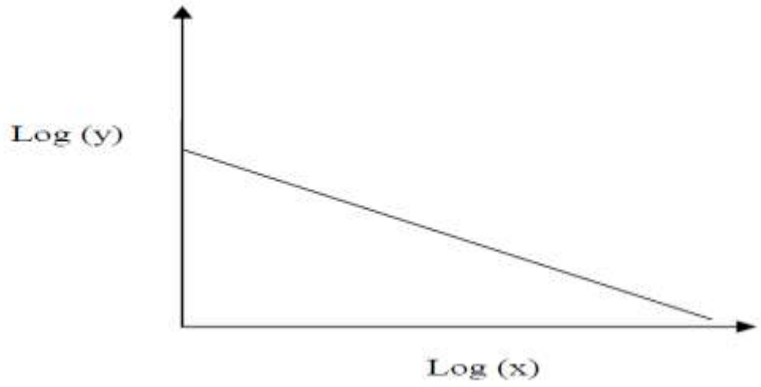

Figure 2: Représentation d'une loi puissance dans un repère bi-logarithmique.

3. Les lois de Zipf et de Zipf inverse

Dans cette section, nous passerons en revue les lois puissance Zipf et Zipf inverse.

3.1. La loi de Zipf

La loi de Zipf désigne une loi empirique énoncée en 1949 [17], décrivant la répartition statistique des fréquences d'apparition des individus d'un ensemble, donnons l'exemple des mots d'un texte.

D'après Zipf, les n-uplets de symboles d'un ensemble organisés topologiquement ne se distribuent pas de façon aléatoire. En effet, si nous classons les n-uplets de symboles selon un ordre décroissant de leurs fréquences d'apparition, nous obtenons la suite (Nσ(1), Nσ(2)… Nσ(n)) des fréquences d'apparition organisées selon une loi puissance. La fréquence d'un n-uplet de rang i vérifie la formule suivante [18]:

$$N_{\sigma(i)} = k.i^{-\alpha} \qquad (II.1)$$

Où k et α sont des constantes. Cette loi puissance est caractérisée par la valeur α de la puissance.

3.2. La loi de Zipf inverse

La loi de Zipf inverse est à son tour une loi empirique décrivant les fréquences d'apparition des n-uplets d'un ensemble. Nous nous intéressons dans ce cas aux nombres de n-uplets ayant la même fréquence d'apparition. Notons que ces n-uplets sont définis exactement de la même façon qu'avec la loi de Zipf.

D'après la loi de Zipf inverse, le nombre I de n-uplets ayant une même fréquence d'apparition f se modélise par la formule suivante :

$$I(f) = A\, f^{-\gamma} \qquad (II.2)$$

Dans cette formule, A et γ sont des constantes positives.

4. Les domaines d'application des lois de Zipf et Zipf inverse

Bien que les lois de Zipf et de Zipf inverse ont été appliquées essentiellement dans la linguistique, nous les rencontrons également dans différents domaines, nous pouvons citer : facteur d'impact des journaux [19], la répartition géographique des différentes populations [20] ainsi que l'analyse d'accès et trafic d'Internet [21].

Quant au domaine du traitement des images, les lois de Zipf ont fait l'objet d'une récente application. En effet, ces lois ont été appliquées avec succès pour la classification des tumeurs mammaires [22], l'extraction des zones d'intérêts ainsi que la détection des objets artificiels au sein de milieux naturels [23], la mesure de la qualité des images compressées [24], la stéganalyse [25].

5. Analyse des images par lois de Zipf et de Zipf inverse

Il a été démontré qu'il existe des répartitions en loi puissance dans le cas bidimensionnel des images [26]. En effet, l'enjeu majeur consiste à la modélisation de la fréquence d'apparition des motifs rencontrés dans l'image selon des distributions en loi de puissance; dans ce sens, ces modèles ont la capacité de caractériser la complexité structurelle des textures présentes dans les images.

Cette analyse est caractérisée par le traçage des courbes de Zipf et de Zipf inverse, ceci dit, ces courbes ne fournissent pas de mesures de texture pouvant être considérées comme descripteurs. L'information de la texture sera quantifiée au moyen d'un ensemble de descripteurs calculés à partir des courbes de Zipf et de Zipf inverse et modélisant l'information sur la distribution spatiale des niveaux de gris dans l'image [22].

5.1. Analyse des images par la loi de Zipf

Le principe d'analyse d' une image par la loi de Zipf [26] consiste à parcourir l'image avec un masque de dimension m*m où tous les motifs rencontrés subiront un codage appliqué aux pixels relatives à l'image.

Par la suite, nous classons ces motifs selon l'ordre des fréquences décroissantes tout en attribuant le rang de chaque fréquence.

La dernière étape consiste à tracer la courbe rang fréquence dans un repère bi-logarithmique.

Nous modélisons cette relation au moyen de la formule suivante [26] :

$$N\sigma(i) = k\, i^{-a} \qquad (II.3)$$

Dans la formule (II.3), $N_{\sigma(i)}$ représente la fréquence d'apparitions d'un motif de rang i, tandis que k et a présentent des constantes.

5.2. Analyse des images par la loi de Zipf inverse

Le principe d'analyse d'une image par la loi de Zipf inverse est le même que celui de la loi de Zipf avec une seule différence que la loi de Zipf inverse s'intéresse aux motifs distincts ayant la même fréquence d'apparition.

D'après la loi de Zipf inverse, le nombre I des motifs différents ayant une fréquence d'apparition f est donné par la formule suivante [26] :

$$I(f) = a\, f^{-\gamma} \qquad (II.4)$$

Dans la formule (II.4), a et γ sont des constantes.

5.3. Codage de l'image

Pour le passage de l'application des lois de Zipf et Zipf inverse du texte à l'image, une adaptation était nécessaire [26]. Les symboles considérés en imageries sont les niveaux de gris qui encodent les pixels. A cet effet, les n-uplets seront considérés sous la forme d'une suite des niveaux de gris constituants des masques qui sont susceptibles d'avoir des formes différentes : nous distinguons des masques carrés 3x3 ou bien des masques linéaires verticaux 3x1, 8x1 ou bien horizontaux, 1x3, 1x8.

Le choix des formes du masque dépend essentiellement du type de motif en vue de recherche, que ce soit un motif linéaire ou bien un motif surfacique.

Evidemment, il est possible d'utiliser d'autres dimensions des motifs, néanmoins une taille plus grande des motifs génèrera le fait que les motifs d'images n'auraient qu'une minime probabilité de réapparaitre plusieurs fois dans l'image. Dans ce sens, la distribution des fréquences des motifs dans l'image n'aurait pas une vraie signification.

Une autre nécessité pour l'analyse des images par des modelés de lois puissance est la définition d'un codage aux motifs de l'image dans le but de réduire le nombre des motifs distincts et augmenter la probabilité qu'un motif se rencontre plusieurs fois au sein de l'image à analyser.

Nous allons introduire les deux codages existants comme suit :

- **Utilisation du codage des 9 classes**

Ce codage vise la réduction du nombre de motifs distincts par la division équitable de l'échelle des niveaux de gris [0,255] en 9 intervalles, ou classes.

Nous numérotons ces intervalles en ordre croissant de 0 à 8 comme mentionné sur le tableau 1

Tableau 1: Partitionnement des niveaux de gris en 9 classes.

0-27	28-55	56-83	84-111	112-139	140-167	168-195	196-223	224-255
0	1	2	3	4	5	6	7	8

Le processus du codage de l'image est comme suit : chaque pixel du motif, aura une codification qui consiste en une valeur de classe c(x, y) obtenue de la valeur g(x, y) de son niveau de gris en se basant sur la formule suivante :

$$c(x,y) = \text{int}\left[\frac{N * g(x,y)}{255}\right] \qquad (II.5)$$

N = 9

Un exemple de ce codage des 9 classes est fourni sur la figure II.3, où nous exposons un motif avant et après l'avoir encodé.

(a)

0	20	250
70	10	213
52	125	120

(b)

0	0	8
2	0	7
1	4	4

Figure 3: Motif original d'une image en (a) et son encodage par la méthode des neufs classes en (b).

- **Utilisation du Codage des rangs généraux**

Il a été prouvé dans [26] [5] que le codage par rang généraux permet d'étudier la texturation fine des images avec l'ultime but de diminuer le nombre de motifs présents dans l'image.

Coder une image par les rangs généraux revient à remplacer les niveaux de gris relatifs aux pixels par leur rang au sein d'un voisinage considéré. Ceci, en ordonnant les niveaux de gris des pixels du motif par un ordre croissant et en affectant la valeur 0 pour le niveau de gris le plus bas. Par la suite, nous incrémentons la valeur d'une unité jusqu'à l'atteinte du niveau de gris le plus élevé.

Mentionnons que les pixels avec la même valeur de niveau de gris auront le même rang.

La figure 4 expose un exemple d'un motif suite à son encodage par la méthode des rangs généraux.

(a)

250	200	200
25	4	29
35	4	35

(b)

5	4	4
1	0	2
3	0	3

Figure 4: Motif original (a) et son encodage avec la méthode des rangs généraux (b).

5.4. Représentation graphique des lois de Zipf et de Zipf inverse

- **Construction de la courbe de Zipf**

L'algorithme de construction de la courbe de Zipf d'une image [26] consiste à :

1- réaliser un balayage séquentiel de l'image sur la base d'un masque de capture de dimension 3x3.
2- Ensuite ; à appliquer une codification des motifs selon le choix d'un codage adéquats aux propriétés que nous souhaitons explorer dans l'image analysée.
3- L'étape suivante consiste à compter le nombre d'occurrences relatifs aux motifs distincts dans l'image. En effet, chaque motif rencontré sera rangé dans un tableau et comparé aux motifs déjà existant dans le tableau, s'il est présent, l'incrémentation de sa fréquence d'apparition d'une unité est appliquée, dans le cas contraire, nous l'indexerons dans le tableau en fixant sa fréquence d'apparition à 1.
4- Après avoir balayé entièrement l'image, les fréquences d'apparition des motifs doivent être triées dans un ordre décroissant.

5- La dernière étape de l'algorithme consiste à tracer la fréquence de chaque motif en fonction de son rang dans un repère bi-logarithmique où l'abscisse représente le rang R des motifs tandis que l'ordonnée représente leur fréquence d'apparition.

Nous présentons sur la figure 5 la courbe de Zipf d'une image.

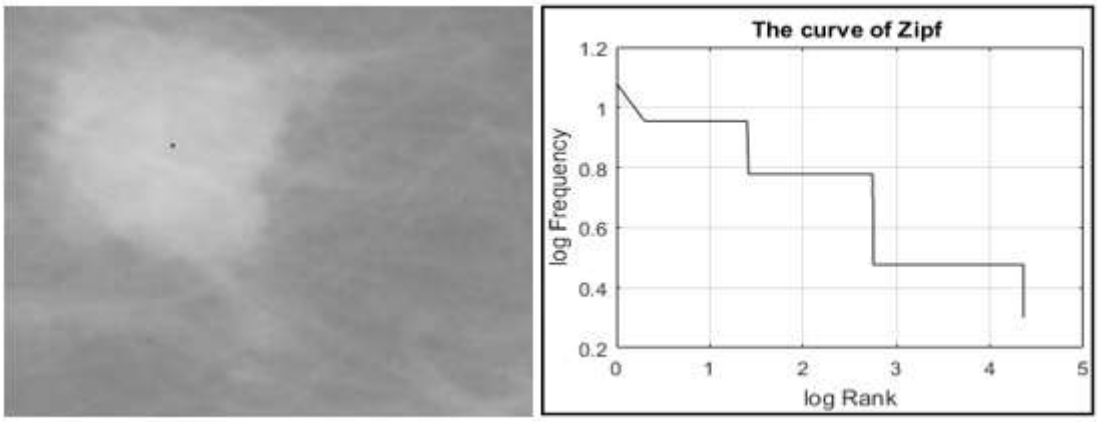

Figure 5:Courbe de Zipf d'une image.

- **Construction de la courbe de Zipf inverse**

L'algorithme de construction de la courbe de Zipf inverse comporte les étapes suivantes [26]:

1- La première étape est identique à celle de la loi de Zipf à savoir de balayer séquentiellement l'image par un masque de capture de taille 3x3, ensuite, les motifs seront encodés par un codage qui correspond aux propriétés de l'image dans le but de dénombrer les motifs trouvés.
2- La deuxième étape consiste à compter les motifs ayant la même fréquence d'apparition que la fréquence courante. Pour ceci, nous initialisons la fréquence cherchée à 1 pour parcourir séquentiellement le tableau des motifs et compter ceux ayant la même fréquence que la fréquence courante.
3- Nous réitérons l'algorithme en incrémentant à chaque fois la fréquence cherchée d'une unité jusqu'à l'atteinte de la fréquence maximale.
4- La dernière étape de l'algorithme consiste à tracer le nombre de motif est en fonction de leur fréquence d'apparition dans un repère bi-logarithmique désignant la courbe de Zipf inverse.

Nous présentons sur la figure 6, la courbe de Zipf inverse d'une image.

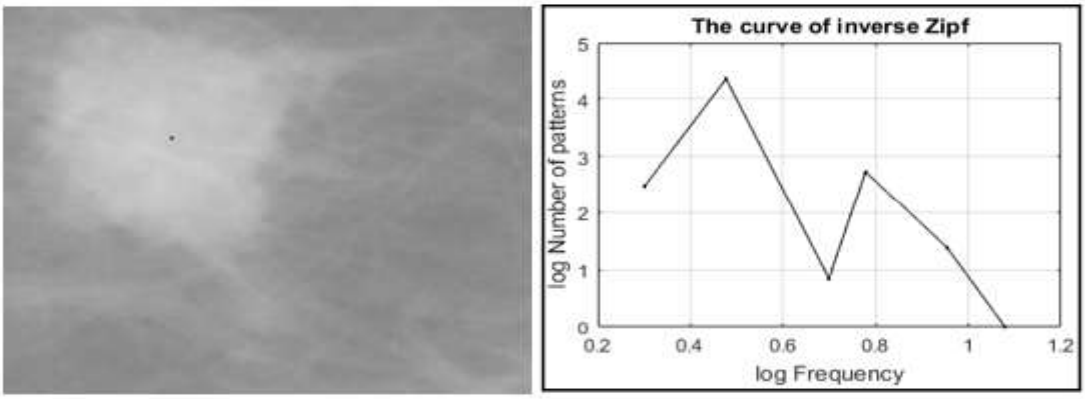

Figure 6:Courbe de Zipf inverse d'une tumeur maligne.

6. Conclusion

Les distributions statistiques en loi puissance sont communes et ont été mises en évidence, initialement, de façon empirique. En effet, elles ont été incorporées dans les domaines très variés : naturels ou bien de sciences humaines.

Dans le domaine de l'analyse d'image et la vision par ordinateur, les lois de Zipf et de Zipf inverse caractérisent la complexité structurelle d'une image en modélisant la répartition statistique des fréquences d'apparition des motifs. Dans ce sens, une nouvelle approche d'analyse de la texture de l'image a été évoquée dans ce mémoire.

Le prochain chapitre sera dédié à l'aide au diagnostic médical du cancer du sein assisté par ordinateur où nous évoquerons en détails un état de l'art de sur l'explosion de l'apprentissage automatique dans ce domaine pour en tirer la problématique de réalisation de ce mémoire.

Chapitre III

L'apprentissage automatique en action : état de l'art sur les approches assistées par ordinateur de pronostic du cancer du sein

1. Introduction

Récemment, des méthodologies d'apprentissage automatique ont été adoptées pour améliorer la précision de détection ainsi que de classification des lésions mammaires, devenant ainsi des éléments clés pour le diagnostic assisté par ordinateur (CAD). En effet, ce domaine est devenu le domaine de recherche le plus actif en imagerie médicale fournissant un deuxième avis pour aider le médecin à donner sa décision pour le diagnostic du cancer du sein.

La motivation principale de l'apprentissage automatique est que l'intelligence et l'apprentissage supplémentaires sont nécessaires pour faire face à des situations incertaines. En effet, les techniques d'apprentissage automatique sont implémentées dans les systèmes CAD et sont avantageuses pour augmenter le taux de survie, car elles aident à automatiser le processus de décision, à fournir une plus grande précision, à répondre immédiatement en cas d'urgence et à minimiser les efforts fournis par les médecins, en particulier lorsqu'il y a une pénurie de personnel médical.

La performance des systèmes d'aides au diagnostic médical assistés par ordinateur (CAD) est souvent reliée à leur puissance de discrimination et de généralisation durant le processus de classification de nouvelles structures.

2. Qu'est-ce que l'apprentissage automatique?

Il n'y a pas de définition universelle de l'apprentissage automatique [27]. Néanmoins, l'apprentissage automatique est généralement considéré comme un domaine de l'intelligence artificielle qui consiste à identifier des modèles à partir de données et à utiliser ces modèles pour faire des prédictions sur des données nouvelles.

Le principal résultat de l'apprentissage automatique est une mesure de généralisation : la mesure dans laquelle le modèle est capable de produire des prédictions correctes avec de nouvelles données, basées sur des règles apprises durant le processus d'apprentissage. En effet, c'est crucial que le modèle apprenne des cas qui sont généralisables pour réaliser des prédictions précises sur les nouvelles données.

Selon [28], l'apprentissage automatique est l'étude d'algorithmes informatiques pour aider à formuler des prédictions et des réactions précises dans certaines circonstances, ou pour agir intelligemment. En général, l'apprentissage automatique consiste à apprendre à créer de meilleures circonstances à l'avenir en fonction de ce qui a été appris dans le passé. Les machines apprennent des informations, des connaissances et de l'expérience existantes; par conséquent, l'apprentissage automatique est le développement de programmes qui nous permettent d'analyser des données provenant de différentes sources en sélectionnant des données pertinentes et en utilisant ces données pour prédire le comportement du système dans des scénarios similaires ou différents.

3. Les différents types d'apprentissage automatique

Nous distinguons plusieurs types d'apprentissage automatique, catégorisés en fonction des méthodes appliquées ainsi que des applications souhaitées [27] :

3.1. Apprentissage supervisé

Dans l'apprentissage supervisé, l'algorithme a accès à ce qu'il essaie de prédire, c'est-à-dire, la variable cible; cela peut être, par exemple, la présence ou absence de maladie ou la gravité des symptômes. Le but ici est d'utiliser un algorithme pour apprendre la fonction optimale qui capture la relation entre l'entrée et la variable cible. La raison pour laquelle ce type d'apprentissage est appelé «supervisé» est que l'algorithme a une connaissance préalable de ce que devraient être les valeurs de sortie (par exemple : tumeur maligne vs tumeur bénigne).

L'algorithme est formé en utilisant plusieurs exemples. Dans ce contexte, l'apprentissage est donc un processus itératif de prédictions et ajustements ultérieurs, jusqu'à ce que la différence entre les prédictions de sortie et la cible soit minimisée autant que possible. La performance est mesurée en comparant les prédictions de l'algorithme aux vraies valeurs cibles dans des données nouvelles. La tâche d'apprentissage supervisé est problème de classification où les algorithmes de classification visent à prédire l'appartenance à une classe, d'un ensemble d'observations ou bien problème de régression. Nous exposons sur la figure suivante une schématisation résumant l'apprentissage supervisé.

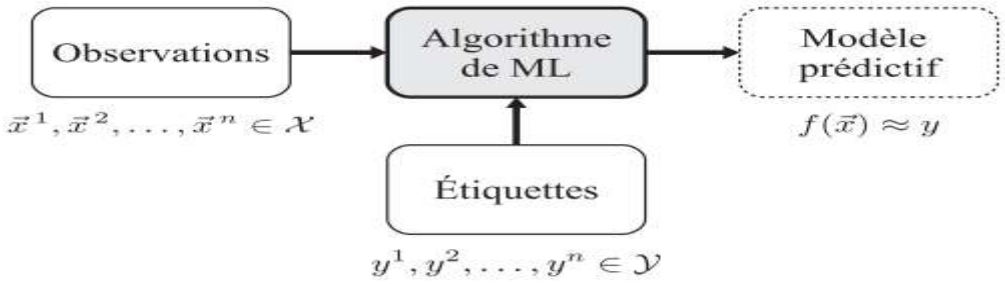

Figure 7: Apprentissage supervisé. Source [29]

3.2. Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, dans une tâche d'apprentissage non supervisé, il n'y a aucune valeur cible. Le but est plutôt de découvrir les structures sous-jacentes dans les données.

La tâche d'apprentissage non supervisé est problème de clustering. En effet, l'analyse de cluster est une technique analytique pour développer des sous-groupes d'un échantillon plus large, tels que des sous-groupes de patients avec différents profils cliniques au sein d'une plus grande cohorte de patients. Ici, les individus sont classés en plus petits groupes non prédéfinis en fonction des similitudes observées entre eux. Nous exposons sur la figure suivante une schématisation résumant l'apprentissage non supervisé.

Figure 8:Apprentissage non supervisé. Source [29]

4. Quelques algorithmes d'apprentissage automatique

4.1. Les machines à vecteurs de support

Les machines à vecteurs de support (SVM : Support Vector Machine) [30] ont été introduites par Vapnik en tant que modèle d'apprentissage automatique basé sur un noyau pour les tâches de classification et de régression. L'extraordinaire capacité de généralisation des SVM, avec sa solution optimale et sa discrimination puissante, a attiré l'attention du data mining, de la reconnaissance des formes et de l'apprentissage automatique au cours des dernières années.

SVM a été utilisé comme un outil puissant pour résoudre des problèmes de classification binaires. Il a été démontré que les SVM sont supérieurs à d'autres méthodes d'apprentissage supervisé. En raison de sa bonne fondation théorique ainsi que bonne capacité de généralisation, les SVM sont devenus l'une des méthodes de classification les plus couramment utilisées.

Les fonctions de décision sont déterminées directement à partir de l'ensemble d'apprentissage en utilisant les SVM de telle sorte que la séparation existante (marge) entre les bordures de décision est maximisée dans un espace hautement dimensionnel appelé espace des descripteurs. Cette stratégie de classification minimise les erreurs de classification des données d'apprentissage et obtient une meilleure capacité de généralisation. Les compétences de classification des SVM et d'autres techniques diffèrent considérablement, en particulier lorsque le nombre de données d'entrée est petit.

Un avantage important des SVM réside dans le fait qu'ils obtiennent un sous-ensemble de vecteurs de support pendant la phase d'apprentissage, qui n'est souvent qu'une petite partie de l'ensemble de données d'origine. Cet ensemble de vecteurs de support représente une tâche de classification donnée et est formé par un petit ensemble de données. Nous présentons sur la figure 9 la modélisation du concept des machines à vecteur de support.

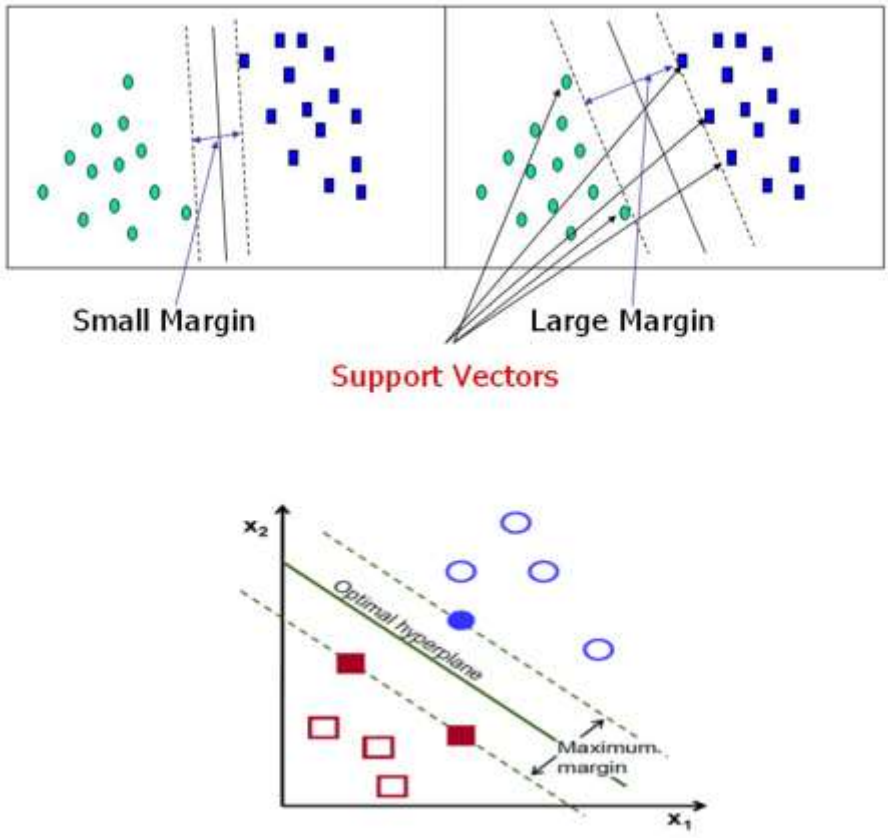

Figure 9: Modélisation du concept des machines à vecteurs de support.

Un SVM est capable de séparer des frontières très complexes où les données ne sont pas linéairement séparables. Ceci, en utilisant une fonction de transformation d'espace (fonction noyau). En effet, Les SVM utilisent divers types de noyaux pour transformer les données non linéairement séparables en données linéairement séparables au biais du changement de l'espace de représentation des données d'entrées en un espace de plus grande dimension, dans lequel il est probable d'exister une séparation linéaire [31] [32]. En général, le noyau RBF est utilisé pour l'apprentissage du classifieur car il est plus efficace et puissant que les deux autres noyaux. Nous présentons sur la figure 10 un exemple d'un problème non linéairement séparable.

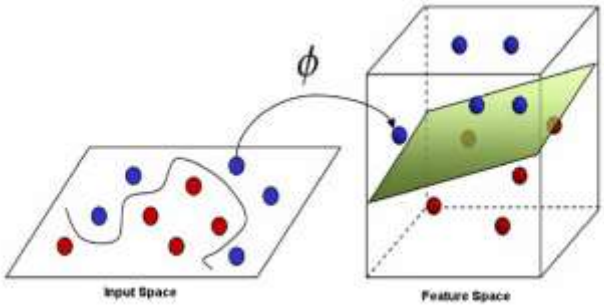

Figure 10:Exemple d'un problème non linéairement séparable où la courbe devient une bande linéaire suite à l'application de la transformation ϕ non linéaire.

4.2. Approches des k plus proches voisins

L'une des premières formes d'utilisation des données étiquetées en imagerie médicale était l'approche des k-plus proches voisins (KNN : K Nearest Neighbors) [33]. Les données d'apprentissage sont projetées sur l'espace des descripteurs, par la suite, un nouvel ensemble d'observations non étiquetées qui vont être classées sur la base d'une combinaison pondérée des K observations étiquetées désignant les plus proches à la nouvelle observation dans l'espace des descripteurs. En effet, La classe de la variable cible, est celle qui est la plus représentée parmi les k plus proches voisins.

4.3. Les arbres de décision

Les arbres de décision appartiennent aux méthodes d'apprentissage supervisés non paramétriques qui sont largement utilisées en classification et en régression. En effet, ceci revient à leur simplicité algorithmique ainsi que leur facilité durant le processus d'interprétation et d'analyse des résultats générés.

Les arbres de décision se fondent sur une approche algorithmique et se visualisent sous forme d'un arbre avec des règles identifiant les techniques de fractionnement d'un ensemble de données. L'ultime enjeu est la création d'un modèle ayant comme objectif la prédiction de la valeur d'une variable cible en se basant sur les règles de décision [34].

5. Application de l'apprentissage automatique à la classification des images médicales

Au long de ces dernières années, il y a eu une accélération du nombre d'articles centrés autour de l'application de l'intelligence artificielle (IA) et plus précisément sa branche consistant en l'apprentissage automatique (ML : Machine Learning) dans le domaine de la santé numérique [35] [36] [37]. Cette augmentation d'intérêt à tirer parti de l'IA et de l'apprentissage automatique dans le domaine de la santé fut suscitée par l'impact de la numérisation et des technologies numériques. À cet égard, la norme de la définition du terme santé numérique s'énonce comme suit : «l'utilisation de l'information et des technologies pour améliorer la santé humaine ».

La croissance de la charge de travail rend le maintien d'une analyse efficace du flux des données médicales, une tâche difficile pour les radiologues et les médecins [38]. En effet, suite aux progrès de l'apprentissage automatique et des techniques de calcul durant ces dernières années, des méthodes informatisées fiables ont été développé pour aider les radiologues et médecins dans le processus d'analyse d'images à différentes étapes du diagnostic et de la prise en charge de la maladie pendant la phase d'élaboration du traitement aux patients.

L'analyse automatisée a été reconnue comme un domaine de recherche important en imagerie médicale [35] et une fois combinée avec les technologies de l'intelligence artificielle et de l'apprentissage automatique, l'analyse de ces données a le potentiel d'identifier de nouveaux schémas à long terme et des facteurs de risque pour améliorer le diagnostic, déclencher des interventions précoces et découvrir des traitements plus efficaces.

De tels paradigmes supervisés de l'apprentissage automatique commencent avec des données [35], par exemple, des images mammaires étiquetées avec un résultat spécifique, tel que l'emplacement d'une tumeur, un algorithme d'apprentissage automatique spécialisé est ensuite utilisé pour relier les observations (qui peuvent ne pas être évidentes pour un être humain) appartenant au flux de données avec le résultat correspondant. Une fois que cet algorithme a identifié le modèle saillant dans les données, l'étape suivante consiste à prédire les résultats, par exemple, l'emplacement d'une tumeur dans une nouvelle mammographie.

Malgré les avantages que l'apprentissage automatique peut apporter à la santé numérique, les défis sont nombreux et doivent être surmontées pour réaliser un tel potentiel pleinement.

Les résultats pertinents en apprentissage automatiques dépendent fortement de la disponibilité de grandes quantités de données de haute qualité représentatives des patients cible. En effet, des ensembles de données peu représentatifs introduisent des biais dans le processus de

l'apprentissage automatique, ce qui à son tour pourrait conduire à des scénarios coûteux d'erreurs ou de overfitting.

6. L'apprentissage automatique pour l'aide au diagnostic médical du cancer du sein assisté par ordinateur

6.1. Le cancer du sein

Le cancer du sein est une cause majeure de décès chez les femmes [39] [40], en particulier dans les pays développés, c'est la deuxième malignité la plus courante. Les cellules malignes peuvent se diviser pour former une nouvelle tumeur dans d'autres parties du corps [40]. Plus précisément, le symptôme le plus courant du cancer du sein est une nouvelle masse. Ce cancer peut également se manifester par des modifications de la taille ou de la forme du sein ou bien de la couleur de la peau du sein, des douleurs au mamelon ou du mamelon tourné vers l'intérieur, des capitons cutanés, des douleurs mammaires et bien d'autres signes.

Bien que les raisons précises de cette maladie ne soient pas encore connues, certains facteurs de risque augmentent la probabilité de développer un cancer du sein, tels que: les facteurs génétiques, l'âge, la consommation de tabac, le surpoids et une ménopause tardive.

À cet égard, plusieurs méthodes peuvent être utilisées pour diagnostiquer cette maladie [40], y compris la biopsie mammaire, l'échographie, la mammographie, la thermographie et la cytologie par aspiration à l'aiguille fine. Cependant et jusqu'à présent, la mammographie s'est avérée être la technique la plus utilisée et la plus fiable pour le dépistage du cancer du sein. Néanmoins, cela n'est parfois pas suffisant et les médecins peuvent avoir besoin d'une biopsie supplémentaire avant de rendre leur dernière décision.

6.2. La mammographie de dépistage du cancer du sein

La mammographie est une technique de radiographie, particulièrement adaptée à l'organe du sein [41]. Elle est conçue pour détecter les anomalies précoces avant qu'elles ne provoquent des symptômes cliniques. La mammographie n'est pas seulement pratiquée dans les campagnes de dépistage du cancer du sein, mais aussi pour le diagnostic ainsi que la localisation lors d'interventions chirurgicales comme le processus de ponctions par exemple. L'intérêt d'un tel examen est qu'il permet d'examiner tout le tissu mammaire.

Lorsqu'une femme ne présente aucun symptôme du cancer du sein, une mammographie de dépistage est effectuée. Ce processus peut réduire le nombre de décès des femmes atteintes d'un cancer du sein à l'âge de 40 à 70 ans [42].

Un radiologue examine les images de la mammographie pour rechercher des signes de tumeurs ou excroissances anormales. Une calcification de petite taille indique un stockage de calcium dans le sein et il est représenté par une tache sur l'image. Jusqu'à 50% des masses présentent des calcifications à l'échelle miniaturisée groupées et dans divers cas, les groupes sont la principale indication du danger [39].

L'interprétation et l'analyse d'images médicales sont l'un des processus les plus difficiles dans des applications avancées de reconnaissance de formes et de vision par ordinateur. En mammographie numérique, la principale difficulté est de faire la distinction entre les tumeurs bénignes et malignes. Dans le cas de la mammographie aux rayons X, la précision du taux de détection est à un niveau faible de l'ordre de 60 à 70% [40] avec un taux d'échec de 4% à 34% [39]. En conséquence, le patient sera confronté à des tests supplémentaires qui peuvent être coûteux et exigeants. Dans ce sens, les chercheurs ont exploité les techniques d'apprentissage automatique (ML) pour faire recours aux limites de ces techniques traditionnelles en aidant les médecins avec un deuxième avis, et donc réduire les erreurs humaines potentielles qui peuvent coûter la vie du patient.

6.3. Détection et classification assistées par ordinateur (CADe/CADx) des tumeurs dans la mammographie

L'application des techniques du traitement d'image et de l'apprentissage automatique a contribué à la tâche de pronostic du cancer, aboutissant à un diagnostic plus précis. Par conséquent, il y a eu un intérêt accru au cours de la dernière décennie pour le développement d'un système de détection et de diagnostic assisté par ordinateur (CADe pour Computer Aided Detection et CADx pour Computer Aided Diagnosis, respectivement) qui donne une aide ou un deuxième avis aux radiologues durant l'interprétation des examens mammographiques [43].

Les systèmes d'aide au diagnostic médical (CAD) se développent avec les méthodes d'apprentissage automatique. En effet, les approches d'apprentissage automatique dans le diagnostic médical assisté par ordinateur (CAD) en imagerie médicale se basent sur les méthodes d'analyse d'image pour reconnaître la maladie et distinguer les différentes classes des structures présentes sur les images, par exemple normales ou anormales, malignes ou bénignes. Les développeurs des CAD conçoivent des technique de traitement d'images ainsi que d'extraction des caractéristiques permettant de distinguer les différentes structures en les

traduisant en valeurs numériques. Dans [44], les auteurs étudient l'état de l'art concernant les systèmes de détection / diagnostic assistés par ordinateur (CAD) pour le cancer du sein en se basant sur des classifieurs appartenant à l'apprentissage automatique. Il est rapporté que la sensibilité de détection sans CAD est d'environ 80% et peut atteindre les 90% suite à l'incorporation des CAD.

6.3.1. Détection des tumeurs assistée par ordinateur (CADe)

Le processus de détection des tumeurs assisté par ordinateur (CADe) est l'étape initiale pour le diagnostic des mammographies. En effet, l'enjeu majeur consiste en la détection et localisation des lésions suspectes relatives au tissu parenchyme mammaire suite à l'utilisation d'un ensemble de descripteurs mathématiques spécifiques [45] [46].

Les algorithmes couramment appliqués dans l'extraction des zones d'intérêt (ROI: region of interest) se basent sur l'analyse des pixels. En effet, les pixels formant un tissu pathologique ont des descripteurs différents comparés aux autres pixels formant le tissu sans anomalie.

Les descripteurs utilisés pour la détection des tumeurs peuvent consister en des valeurs de gris ou bien des mesures calculées sur le tissu.

Une fois que les lésions mammaires sont détectées, elles sont introduites dans l'étape de classification basée sur l'apprentissage automatique du système d'aide au diagnostic médical du cancer du sein (CAD) [47].

6.3.2. Diagnostic des tumeurs assisté par ordinateur (CADx)

Le processus de diagnostic des tumeurs assisté par ordinateur (CADx) réalise la caractérisation des lésions détectées par le radiologue ou bien par les CADe. La sortie de ces systèmes est la classification des lésions en malignes / bénignes et dans d'autres cas malignes / bénignes / tissu parenchyme mammaire sein [45] [46].

Les systèmes de diagnostic assisté par ordinateur (CADx) réalisent l'extraction de descripteurs discriminants sur les régions suspectes segmentées. Les descripteurs extraits sont ensuite utilisés comme variables prédictives d'entrée à un classifieur, et un modèle prédictif est formé en ajustant les poids des différents descripteurs se basant sur les propriétés statistiques d'un ensemble d'apprentissage pour estimer la probabilité qu'une zone d'intérêt appartienne à l'une des classes.

6.4. Etat de l'art sur les systèmes d'aide au diagnostic médical du cancer du sein (CAD)

Nous distinguons divers travaux connexes dans la littérature, où des systèmes de diagnostic assisté par ordinateur (CAD) ont été développés.

En effet, différent systèmes CAD commercialisés sont disponibles pour l'analyse des mammographies et sont cliniquement conformes et acceptables, citons AccuDetect Parascript® software for mammography [48], ainsi que R2 ImageChecker or iCAD Second Look [49]. Cependant, il a été démontré que tous les systèmes CAD souffrent d'une précision limitée [50].

Dans [51], les auteurs présentent les résultats liés à la base de données DDSM utilisant comme descripteurs les moments générés par la transformation en curvelette. La précision de détection des anomalies a passé de 81,26% à 86,46%.

Dans [52], les auteurs effectuent la classification des mammographies à l'aide de descripteurs extraits suite à l'application de la transformée de Hough. La validation de l'approche proposée était réalisée sur 95 images mammaires par le classifieur des machines à vecteur de support (SVM) où une précision de 94% a été atteinte.

Dans [53], un modèle amélioré d'aide au diagnostic médical du cancer du sein (CAD) est proposé pour la classification des masses mammaires se basant sur la transformation en ondelettes pour extraire les descripteurs de la région d'intérêt des images mammaires. La dimension des vecteurs de caractéristiques est ensuite réduite en utilisant une fusion de méthodes PCA et LDA. Enfin, la classification est effectuée basée techniques de l'apprentissage automatique. Le modèle a été évalué sur deux ensembles d'images mammaires, à savoir MIAS et DDSM. Suite aux expériences réalisées, le système de CAD proposé obtient des résultats idéaux pour l'ensemble de données MIAS en atteignant une précision de 99,76% (normal vs anormal) et 98,80% (bénin vs malin) pour l'ensemble de données DDSM.

Dans [43], les auteurs utilisent les motifs binaires locaux (LBP: Local Binary Patterns) pour la caractérisation de la texture des régions d'intérêt (ROIs). Sur cette représentation, d'autres représentations ont été générées utilisant des techniques tels que les histogrammes d'image, les matrices de cooccurrence de niveau de gris (GLCM) et la matrice de la longueur de plage de niveau de gris (GLRLM). Une précision de la classification de l'ordre de 88.3% a été atteinte.

7. Conclusion

L'apprentissage automatique est un domaine émergent de l'intelligence artificielle qui prend de l'ampleur dans la recherche sur les tumeurs mammaires. En effet, ce domaine nécessite un grand nombre d'observations et ce défi est maintenant relevé par un nombre croissant d'initiatives de sauvegarde et de stockage de données médicales où l'enjeu majeur consiste à identifier des modèles à partir de ces données et utiliser ces modèles pour faire des prédictions sur des données nouvelles et non explorées.

L'apprentissage automatique est un domaine en pleine croissance avec une multitude de méthodes pour en choisir parmi. En effet, le chapitre prochain illustre la problématique tirée dans ce domaine ainsi que le cadre méthodologique d'apprentissage automatique suivi pour la réalisation de l'approche proposée.

Chapitre VI

Problématique, approche proposée et évaluation des résultats obtenus

1. Introduction

Le domaine de l'intelligence artificielle s'intéresse à la théorie et mise en œuvre de systèmes informatiques capables d'exécuter des tâches qui nécessitent généralement une intelligence humaine [27]. Ceci suite à l'intersection de plusieurs disciplines telles que l'informatique, l'ingénierie, les mathématiques, les statistiques, la psychologie ainsi que la neuroscience.

L'apprentissage automatique est un domaine de l'intelligence artificielle qui a émergé dans le cadre de la quête visant la construction de machines intelligentes capables d'apprendre.

L'ultime but est la création de modèles prédictifs par les algorithmes d'apprentissage supervisé qui se basent sur des informations d'entrées et des sorties connues pour la prédiction de future sorties [31]. Il est utile de mentionner qu'il n'y a pas d'algorithme d'apprentissage automatique adaptés à toutes les exigences des applications liés à la science médicinale. Dans ce sens, nous allons présenter dans ce chapitre la problématique tirée dans cet intéressant domaine d'actualité ainsi que l'approche proposée en indiquant les diverses étapes suivies pour son accomplissement.

2. Problématique

L'apprentissage automatique, en tant que branche de l'intelligence artificielle, offre une myriade d'opportunités pour booster le diagnostic médical et le domaine de la médecine en général. Ceci, par l'analyse objective et approfondie des données médicales vers l'identification de nouveaux modèles à long terme et les facteurs de risque pour faciliter le diagnostic. Plus loin, l'utilisation de l'IA dans les systèmes de santé commercialisables peut offrir une solution vers des traitements adéquats à chaque patient [35].

Pour l'analyse d'images mammaires, l'extraction de descripteurs est l'étape la plus cruciale lors de l'application des techniques d'apprentissage automatique [54]. En effet, la détection des contours des tumeurs malignes et bénignes est très importante car selon les radiologues plus les bordures d'une tumeur sont irrégulières, plus l'association à une tumeur maligne a été observé [55]. Cependant, les cellules malignes sont appelées cellules cancéreuses et deviennent plus dangereuses quand elles commencent à se répliquer ou à métastaser dans d'autres organes du corps. Les cellules bénignes ont une forme bien définie et sont de grande taille alors que la taille des cellules malignes est très petite. En raison de sa petite taille et de la présence de tissus adipeux et denses, il devient très difficile de détecter une tumeur maligne au stage primaire. Par conséquent, des systèmes automatisés ou informatisés avancés sont nécessaires pour détecter la tumeur du sein [56].

Différents descripteurs de forme et de texture sont proposés pour classer les masses détectées dans les images mammaires comme malignes ou bénignes. Une bonne classification est directement liée aux caractéristiques des descripteurs et à la capacité de discriminer les images dans ces deux classes. Ainsi, les chercheurs sont toujours à la recherche de descripteurs faciles à calculer et qui ont le pouvoir de discriminer les classes, contribuant considérablement à la performance de classification [43]. Dans ce sens, notre problématique de recherche est fondée sur un ultime but de contribuer une technique robuste d'analyse et de caractérisation de la texture au sein des images mammaires. Effectivement, les relations linéaires sont connues pour leur simplicité d'application dans divers domaines de recherche, néanmoins, elles représentent un inconvénient majeure celui de leur limites pour la modélisation d'une image avec sa structure complexe.

Les lois de puissance Zipf et Zipf inverse ont été appliquées dans un processus d'aide au diagnostic médical du cancer du sein assisté par ordinateur pour la caractérisation de la texture des images mammaires. Dans ce sens, ces lois ont généré des descripteurs texturaux

discriminants durant la distinction entre les textures mammaires d'un tissu portant une tumeur maligne et celui portant une tumeur bénigne [5]. En effet, nous trions une variante des perspectives soulignées dans ces travaux en proposant une fusion des lois de Zipf et de Zipf inverse avec les filtres de Gabor qui se sont avérés des banques de filtres adaptées pour extraire les caractéristiques texturales de type biologique [57] [58], pour bénéficier de l'apport complémentaire de ces deux approches d'analyse et de caractérisation de la texture. Dans [59] [60] [61] [39], une fusion de diverses approches de caractérisation de la texture a été suggéré où de nettes améliorations de la précision de classification ont été observé suite à cette fusion de plusieurs descripteurs.

La valeur de gris est l'information de base de l'image en niveaux de gris et décrit chaque valeur de pixel. En effet, les différents grades pathologiques du cancer du sein ont des représentations différentes en niveaux de gris dans la région cancéreuse. L'analyse statistique de la valeur de gris dans la région de la tumeur du sein reflète la distribution des intensités de pixels de la tumeur dans l'image, et pourrait fournir une référence à une discrimination. Dans ce sens, nous avons souligné notre motivation aux caractéristiques de texture extraites, jugées bénéfiques pour reconnaître la structure tumorale de différents grades pathologiques grâce à la reconnaissance de texture [9]. En effet, dû à certaines limitations intrinsèques des images mammaires, les descripteurs géométriques ne sont pas puissants pour une discrimination précise entre les tumeurs bénignes et celles malignes. Mentionnons qu'une classification précise des anomalies devient difficile pour les radiologues experts ou bien pour les algorithmes de CAD lorsque les lésions sont masquées ou non spécifiques sur les mammographies. Cette nature de lésions suspectes peut conduire à de faux positifs (FP).

Ainsi, une minimisation des erreurs de diagnostic ainsi qu'une amélioration de la précision devient une tâche cruciale dans le processus de classification. Un remède à cette tâche pourra être acquis à l'aide des descripteurs texturaux vu leur capacité à discriminer les lésions mammaires [62].

Donc, l'enjeu majeur de notre problématique réside dans la caractérisation de la texture, à la fois dans le domaine spatial (les lois de Zipf et de Zipf inverse) et dans le domaine fréquentiel (les filtres de Gabor). En effet, les approches statistiques utilisent les propriétés qui régissent la distribution et les relations des niveaux de gris dans l'image. D'autre part, les méthodes basées transformations réalisent le traitement de l'image dans le domaine de transformation pour extraire les caractéristiques de texture à différentes orientations [63].

L'interprétation et l'analyse d'images mammaires sont l'un des processus les plus difficiles dans des applications avancées de reconnaissance de formes et de vision par ordinateur. En effet, la principale difficulté est de faire la distinction entre les tumeurs bénignes et malignes, le tissu normal et la tumeur. Les descripteurs utilisés seront principalement normalisés pour ensuite les incorporer à la catégorisation des tumeurs bénignes ou malignes ou bien tissu normal ou tumeur via un système CAD qui servira comme un deuxième outil d'avis aux radiologues pour l'analyse des images mammaires basée texture. En effet, notre objectif est de démontrer que ces outils peuvent s'avérer utiles pour améliorer la précision du diagnostic en mettant en évidence les zones de suspicion qui peuvent être manquées lors de l'analyse visuelle [64].

3. Approche proposée basée fusion des lois de puissance : Zipf, Zipf inverse et les filtres de Gabor via un pipeline d'apprentissage automatique pour la classification des tumeurs mammaires

Le réglage des hyperparamètres est une étape cruciale dans toutes les applications d'apprentissage automatique pour assurer la généralisabilité du modèle produit : le modèle fonctionne bien sur les instances de données non utilisées dans la phase d'apprentissage.

Le rôle de la phase de test est d'évaluer cet aspect par l'utilisation d'instances de données jamais traitées auparavant [35]. Nous présentons sur la figure 11 le pipeline d'apprentissage automatique suivi pour l'élaboration de notre approche.

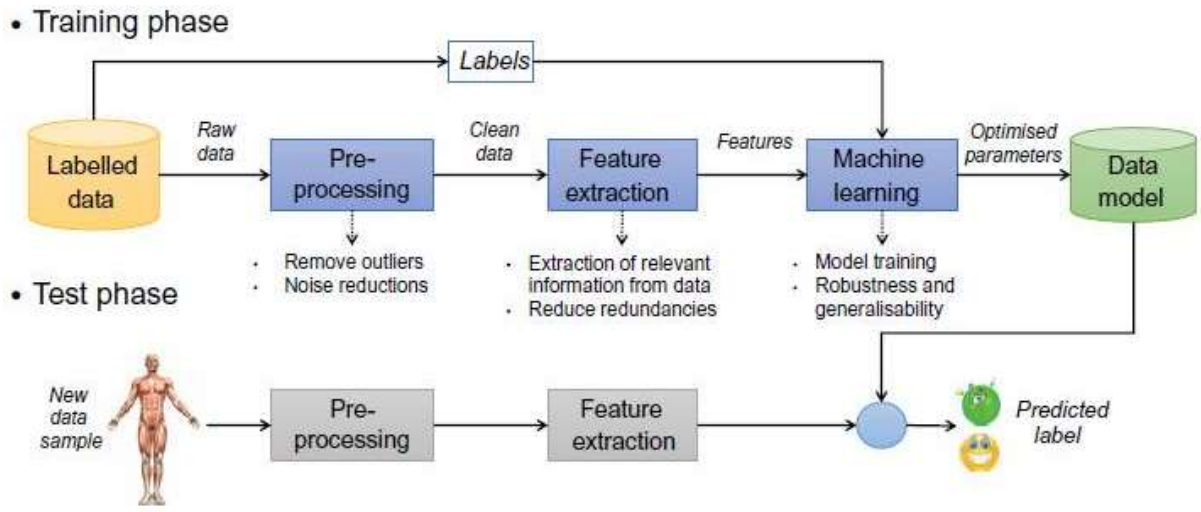

Figure 11:Pipeline d'apprentissage automatique suivi pour l'élaboration de notre approche. Source [35].

3.1. Collecte de données

L'une des exigences de base pour développer un algorithme robuste d'apprentissage automatique est la disposition d'un ensemble d'apprentissage suffisamment grand [38].

En effet, les différentes classes dans les bases d'images devraient idéalement être équilibrées. Si nous considérons le cancer du sein, plusieurs cas par mille dans la population de dépistage sont atteints de ce cancer. A cet effet, il est difficile de recueillir suffisamment de mammographies de dépistage.

Dans le présent travail, nous avons utilisé la base d'images mammaires DDSM [65] qui a été collecté par l'équipe d'experts de l'Université de Floride du Sud. Elle dispose de 2620 cas de l'organe du sein des patientes classés en 43 volumes différents. Pour chaque cas, quatre mammographies ont été recueillies avec deux vues de diapositives différentes pour chaque sein (MLO et CC). Cette base de données est généré avec une taille moyenne de 3 000 × 4 800 pixels, une profondeur de 16 bits et une résolution spatiale de 42 microns. Chaque image a ses propres informations déterminées par des experts, y compris le type de cancer (bénin ou malin) et la localisation de la lésion.

3.2. VI.3.2 Prétraitement

Nous avons réalisé l'extraction des régions d'intérêt à partir des mammographies par l'approche de segmentation présentée dans [5] où les régions d'intérêt ont la dimension de 227x227 pixels et ont été étiquetées comme tissu bénin, tissu cancéreux ou tissu normal en référence aux données accompagnants les mammographies dans la base DDSM. Notons que les régions sans la présence de masses ont été extraites manuellement à partir d'emplacements aléatoires des mammographies appartenants aux cas normaux de cette base de données [66]. Cette procédure d'extraction de régions seines de forme rectangulaire n'affecte pas le résultat car les index décrivent uniquement les informations de texture et non la forme.

Différentes modalités d'imagerie sont utilisées pour analyser les organes du corps, mais le principal problème survient lors de l'acquisition d'image tel qu'un mauvais contraste ou l'intégration du bruit ainsi que la variabilité de l'apparence, de la forme et de l'emplacement de la région anormale. En effet, ces facteurs dégradent considérablement les performances des algorithmes [41].

Des techniques d'amélioration du contraste telles que l'égalisation adaptative de l'histogramme, le filtrage non linéaire, sont appliquées sur la région du sein pour améliorer la visualisation des tissus ou d'une tumeur dans un patch de mammographie [57]. Dans ce sens, l'égalisation d'histogramme est l'une des techniques les plus élémentaires ici, qui étire le contraste des régions d'histogramme élevées et compresse le contraste des régions d'histogramme bas [57]. En conséquence, si la région d'intérêt dans une image n'occupe qu'une petite partie, elle ne sera pas correctement améliorée pendant l'égalisation d'histogramme. L'aspect principal du prétraitement est de mieux capturer la texture de l'organe du sein dans les mammographies.

Dans ce travail nous avons appliqué l'égalisation adaptative de l'histogramme pour l'amélioration du contraste des régions d'intérêt [63]

Nous présentons sur la figure 12 ; deux exemples de régions d'intérêt prétraitées.

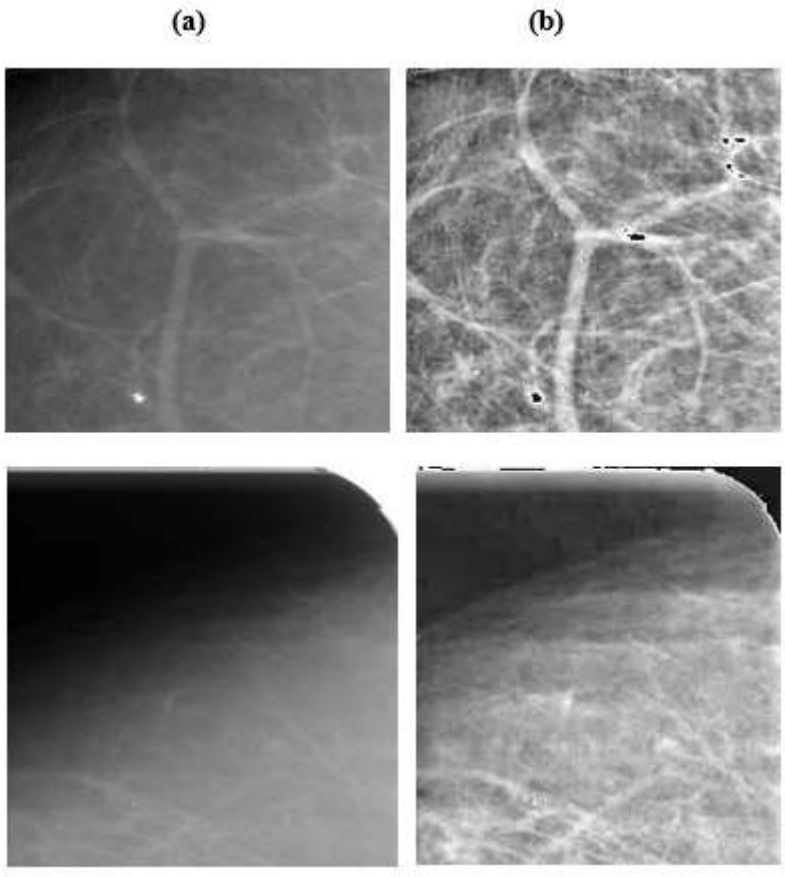

Figure 12:Prétraitement basé égalisation adaptative de l'histogramme : en (a) les régions d'intérêt originales et en (b) les régions d'intérêt prétraitées.

3.3. Extraction des descripteurs

3.3.1. Enjeu majeur de la fusion des aprroches des lois puissance Zipf et Zipf inverse avec les filtres de Gabor pour la caractérisation de la texture

L'étape suivante de la phase d'apprentissage est l'extraction des descripteurs désignant les représentations des données fournies à un algorithme d'apprentissage automatique [35]. Essentiellement, les descripteurs d'images sont la seule information que l'algorithme d'apprentissage automatique utilise pour prendre sa décision.

L'idée de la paramétrisation des descripteurs est de convertir les données brutes dans une représentation avec moins de redondances (informations peu requises et potentiellement de confusion), jugée plus appropriée pour les algorithmes d'apprentissage automatique.

La différence entre le tissu normal et le tissu cancéreux est très faible dans certains cas. Donc la classification des images histopathologiques et l'identification des zones cancéreuses sont assez difficiles en raison de la complexité et de la résolution de l'arrière-plan de l'image [67]. En effet, le choix des descripteurs à utiliser varie en fonction du problème ; donc, vu que nous travaillons sur l'analyse des mammographies et la vision par ordinateur où l'ultime but est d'imiter la vision humaine durant l'élaboration de la tâche du diagnostic médical, nous avons choisi des approches de caractérisation du tissu mammaire qui modélisent la vision humaine de bas niveau. En effet, il a été prouvé que l'attention visuelle est souvent inconsciemment motivée par des niveaux de stimulation faibles de la scène comme le contraste ou l'intensité ou bien la texture [5]. Dans ce sens, les lois de Zipf et de Zipf inverse sont des méthodes perceptuelles basées sur l'analyse du contenu structurel relatif aux images pour la caractérisation de la texture. En effet, nous avons constaté dans [5] que l'humain a tendance à fixer rarement le regard aux régions ayant une distribution gaussienne, en contre partie, il s'intéresse à des régions suivant des distributions exponentielles et le plus souvent aux régions présentant des distributions suivant des lois puissance. D'autres parts, les filtres de Gabor sont conçus pour ressembler aux performances des cellules corticales visuelles du mammifère, dans un sens d'extraction de traits à différentes orientations et échelles [68]. En effet, il a été affirmé que leur fonctionnement est assimilé au fonctionnement de certains neurones du cortex visuel humain par la modélisation de la sensibilité fréquentielle et directionnelle caractéristique du fonctionnement du cortex visuel humain [69] [70].

L'idée de fusionner ces deux approches d'analyse de la texture était basée aussi sur le fait que les filtres de Gabor sont des techniques multi résolutions qui peuvent décomposer la complexité statistique des textures [68]. De plus, leur grande sensibilité aux caractéristiques locales facilite

considérablement les processus de prétention ou subtile discrimination de texture. Les rendent d'excellents extracteurs de textures [10] vu qu'ils sont assez facile à créer et manipuler en détriment de leur puissance.

3.3.2. Analyse et caractérisation de la texture des zones d'intérêt par les lois de Zipf et de Zipf inverse

Les lois de Zipf et de Zipf inverse caractérisent la complexité structurelle relative à la texture d'une image en quantifiant la structure sous-jacente des tissus des lésions mammaires par rapport à leur nature statistique.

Le codage des 9 classes permet de mettre en évidence les principales structures de l'image tandis que le codage des rangs généraux étudie la texturation fine de l'image d'où son adéquation à notre approche proposée.

Le codage des rangs généraux (mentionné dans la sous-section II.5.3) est appliqué sur les pixels des zones d'intérêt pour obtenir les courbes de Zipf et Zipf inverse, mentionnée sur la figure 13et 14 respectivement où nous exposons les courbes de Zipf et de Zipf inverse obtenues à partir de deux zones d'intérêts : l'une portant une tumeur bénigne et l'autre portant une tumeur maligne.

L'utilisation de toutes les intensités de pixels dans les images de la mammographie lors de la classification pouvant contenir de considérables redondances. En effet, nous allons extraire des descripteurs texturaux à partir des courbes de Zipf et Zipf inverse pour les fournir à l'algorithme de l'apprentissage automatique dans le but d'aboutir à de meilleures performances.

- **Analyse des courbes de Zipf relatives aux zones d'intérêt portant une tumeur maligne et une tumeur bénigne**

Dans le cas de la zone d'intérêt portant une tumeur bénigne, nous notons une ordonnée à l'origine élevée du fait que le tissu mammaire est plutôt homogène et nous distinguons ainsi un motif répété plusieurs fois. Par ailleurs, pour la zone d'intérêt présentant une tumeur maligne, nous observons une ordonnée à l'origine de la courbe de Zipf nettement plus basse du fait que la tumeur maligne présente une texture hétérogène causée par la nature invasive des tumeurs malignes et par conséquent nous ne distinguons pas de sur-représentation du motif homogène, bien au contraire, plusieurs motifs hétérogènes provoquent que l'ordonnée à l'origine soit basse.

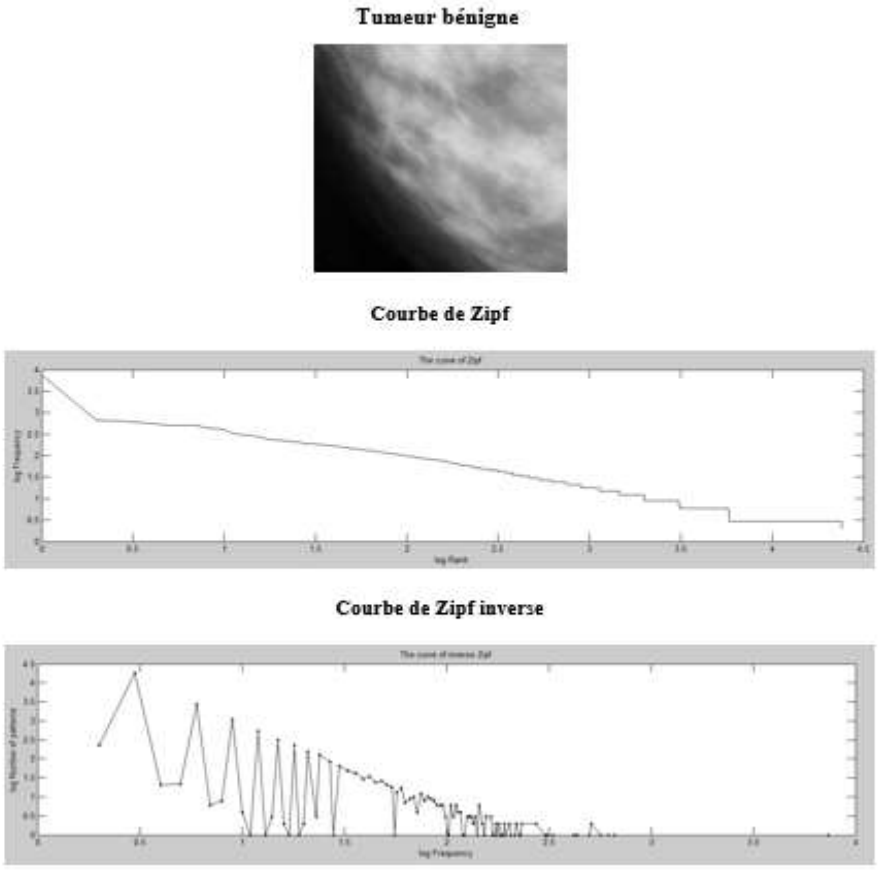

Figure 13:Courbes de Zipf et de Zipf inverse d'une zone d'intérêt portant une tumeur bénigne.

- **Analyse des courbes de Zipf inverse relatives aux zones d'intérêt portant une tumeur maligne et une tumeur bénigne**

Nous pouvons affirmer que la distinction entre les courbes de Zipf inverses pour la zone d'intérêt portant une tumeur bénigne et la zone d'intérêt portant une tumeur maligne est évidente. En effet, l'ordonnée à l'origine de la courbe de Zipf inverse de la tumeur bénigne est inférieure à celle de la courbe de Zipf inverse de la tumeur maligne, ceci revient au fait que la région d'intérêt de la tumeur maligne présente une texture complexe, donc le nombre de l'occurrence est très élevée, ce qui fait que la courbe de Zipf inverse comporte un grand nombre de motifs qui apparaissent une fois.

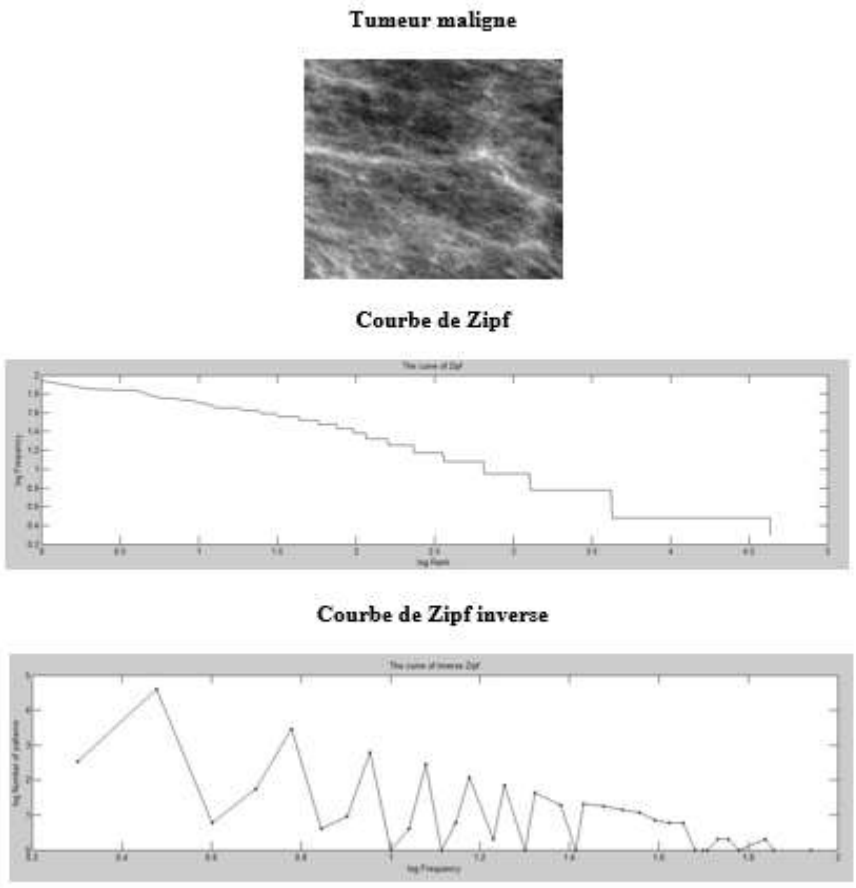

Figure 14:Courbes de Zipf et de Zipf inverse d'une zone d'intérêt portant une tumeur maligne.

Les cellules des tissus cancéreux ont tendance à se dilater et la couleur des tissus devient plus claire. Des irrégularités dans les textures des cellules se sont concrétisées. Dans les tissus normaux ou portants une tumeur bénigne, les textures cellulaires sont plus régulières et la couleur est plus foncée. Suite à ce qui a été analysé en haut, nous pouvons affirmer que l'analyse des zones d'intérêt par les lois de Zipf et de Zipf inverse par un codage des images mammaires à base de rangs, permet de distinguer de manière significative les tumeurs malignes ou bénignes à l'aide des descripteurs calculés et extraits à partir des courbes de Zipf et de Zipf inverses pour chaque zone d'intérêt.

Les descripteurs dérivés des courbes de Zipf et de Zipf inverses sont les suivants [5] :

- **Les pentes des courbes de Zipf et de Zipf inverse**

La pente moyenne d'une courbe est le coefficient directeur de la droite des moindres carrés. Elle est donnée par la formule VI.1 :

$$p = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad \text{(VI.1)}$$

- **L'aire délimitée par la courbe de Zipf**

Nous calculons l'aire délimitée par la courbe de Zipf à partir des courbes de Zipf obtenues suite au codage de l'image par le codage des rangs généraux. Soit *n* le nombre de motifs de la courbe, f_i la fréquence et r_i le rang du motif i, l'aire de la courbe est donnée par la formule VI.2 :

$$A = \sum_{i=1}^{n-1} \frac{(f_i + f_{i+1})(r_{i+1} - r_i)}{2} \qquad \text{(VI.2)}$$

- **Entropie1 de la courbe de Zipf**

L'entropie relative aux motifs des images mammographies est définie par la formule VI.3:

$$H_w = -\sum_{r=1}^{R} \frac{f(r)}{T} \log_R \frac{f(r)}{T} \qquad \text{(VI.3)}$$

Dans cette formule, f (r) représente la fréquence du motif pour la ligne r, T représente le nombre total de motifs différents, et nous utilisons un logarithme avec la base R.

- **Entropie2 de la courbe de Zipf**

L'entropie relative à la fréquence d'apparition des motifs est définie par la formule suivante:

$$H_f = -\sum_{f=1}^{F} \frac{I(f)}{R} \log_F \frac{I(f)}{R} \qquad \text{(VI.4)}$$

Dans cette formule, I(f) représente le nombre de motifs distincts ayant une fréquence d'apparition égale à f et F représente le nombre entier d'occurrences des motifs dans l'image.

- **Les ordonnées à l'origine des courbes de Zipf et Zipf inverse**
- **La constante alpha de la courbe de Zipf**

La loi de Zipf est fortement exprimée de la façon suivante: Quel que soit un motif appartenant à une image, la fréquence d'apparition de ce motif * son rang dans une liste ordonnée décroissante des fréquences d'apparition des motifs = constante [6].

3.3.3. Analyse et caractérisation de la texture des zones d'intérêt par les filtres de Gabor

Les filtres de Gabor effectuent une décomposition multi-résolution vu leur localisation dans le domaine spatial ainsi que l'espace-fréquence d'où leur puissance dans l'analyse de la texture nécessitant à la fois des mesures dans le domaine spatiale et celui spatial-fréquence [10].

Dans le cas 2D, le filtre de Gabor est généré par une gaussienne bidimensionnelle qui est modulée par une fonction sinusoïdale complexe [69]. L'expression du filtre de Gabor 2D est donnée par [69]:

$$h(x,y) = \frac{1}{\sigma x\, \sigma y 2\pi} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \exp\left(j2\pi u_0 (x\cos\theta + y\sin\theta)\right) \quad \text{(VI.5)}$$

L'enveloppe Gaussienne :

$$g(x,y) = \frac{1}{\sigma x\, \sigma y 2\pi} \exp\left(-\frac{x^2}{2\sigma x^2} \frac{y^2}{2\sigma y^2}\right) \quad \text{(VI.6)}$$

Avec : $X = x \cos\theta + y \sin\theta$

$Y = -x \sin\theta + y \cos\theta$

Où σx et σy sont les écarts-types respectivement le long des axes x et y (ou les constants spatiales du filtre), elles déterminent la largeur du filtre.

U_0 : la fréquence centrale.

θ : l'angle de rotation de [x, y] par rapport à (x, y), il donne l'orientation de l'enveloppe Gaussienne g(x, y).

Les filtres de Gabor sont des opérateurs particulièrement commodes pour la caractérisation de la texture. En effet, ils sont capables d'isoler dans une image ; des composantes très variées qui vont de gros objets clairement définis à de fins détails d'orientation particulière, en changeant simplement deux paramètres : la fréquence et l'orientation [11].

Une caractéristique importante des filtres de Gabor est qu'ils peuvent être réglés avec différentes fréquences et orientations [71] pour détecter les changements progressifs de texture et les variations de texture. En effet, les filtres de Gabor sont un type de filtrage passe-bande permettant la conservation de l'information sélectionnée sur le spectre d'une image via la bande de fréquence sélectionnée [11].Ils explorent des propriétés locales dans une région d'intérêt de

l'image. En effet, la configuration classique d'extraction de données spatio-fréquentielles à partir des textures est élaborée par la convolution de l'image avec un banc de filtres [69] où chacun est centré sur une fréquence ainsi qu'une orientation pour couvrir au mieux l'intégralité du domaine fréquentiel [70]. Dans ce sens, chaque pixel appartenant à l'image donnera une réponse à chaque filtre.

Pour des fins de classification, nous examinons la relation existante entre les réponses générées des filtres de Gabor relatives à diverses images.

Nous présentons dans ce qui suit, un exemple de la façon dont les filtres de Gabor peuvent être appliqués sur la tumeur maligne de la figure 15, où les réponses peuvent être utilisées par la suite pour la l'extraction de descripteurs.

Les paramètres les plus importants du filtre de Gabor sont la fréquence radiale et l'orientation. Ils définissent la localisation du canal dans le plan fréquentiel. En effet, chaque image de taille N*N aura des fréquences significatives dans l'intervalle [0..N/4] et à la puissance 2.

Nous tenons à mentionner que nous avons utilisé une banque ou un tableau des filtres de Gabor de taille 39 x 39 selon 5 fréquences et 8 orientations. Comme mentionné sur la figure 15 :

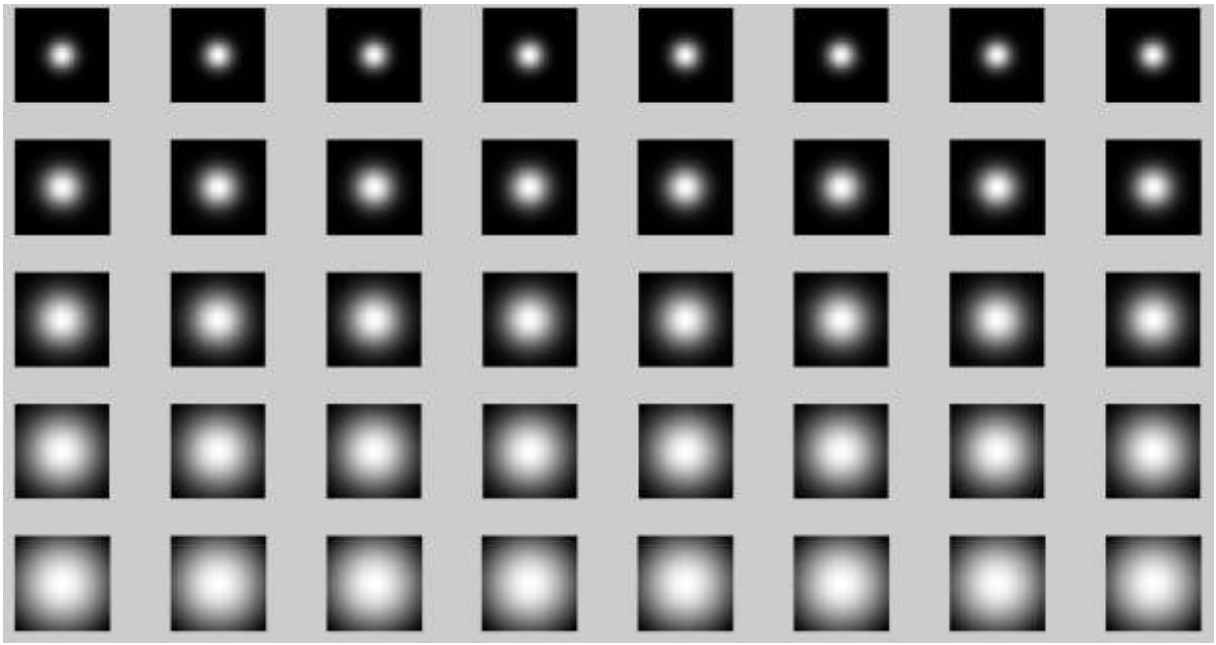

Figure 15:Les filtres de Gabor utilisés dans le domaine fréquentiel.

Ainsi l'utilisation d'un ensemble varié de filtres de Gabor permettra une couverture plus large de l'espace fréquentiel en détectant davantage d'orientations, donc d'extraire tous les contours de l'image [11].

Sur la figure 16, nous exposons la partie réelle de ces filtres.

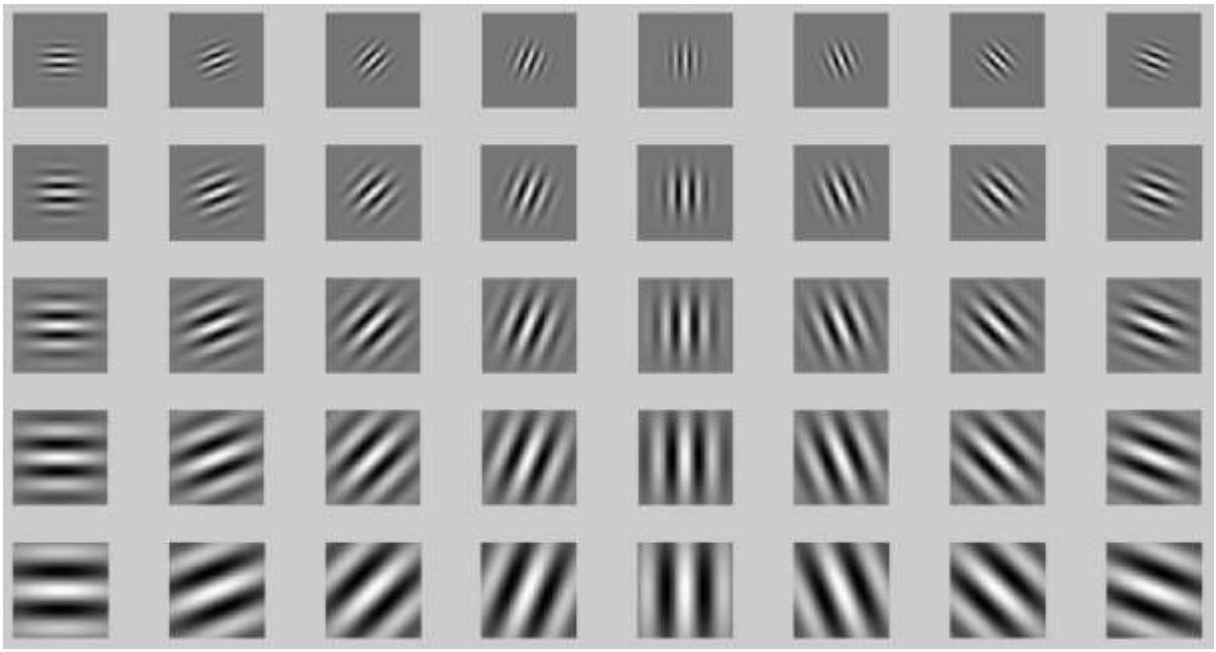

Figure 16:Parties réelles des filtres utilisés

Maintenant, nous allons exposer sur la figure 17, la partie réelle des régions d'intérêt filtrées:

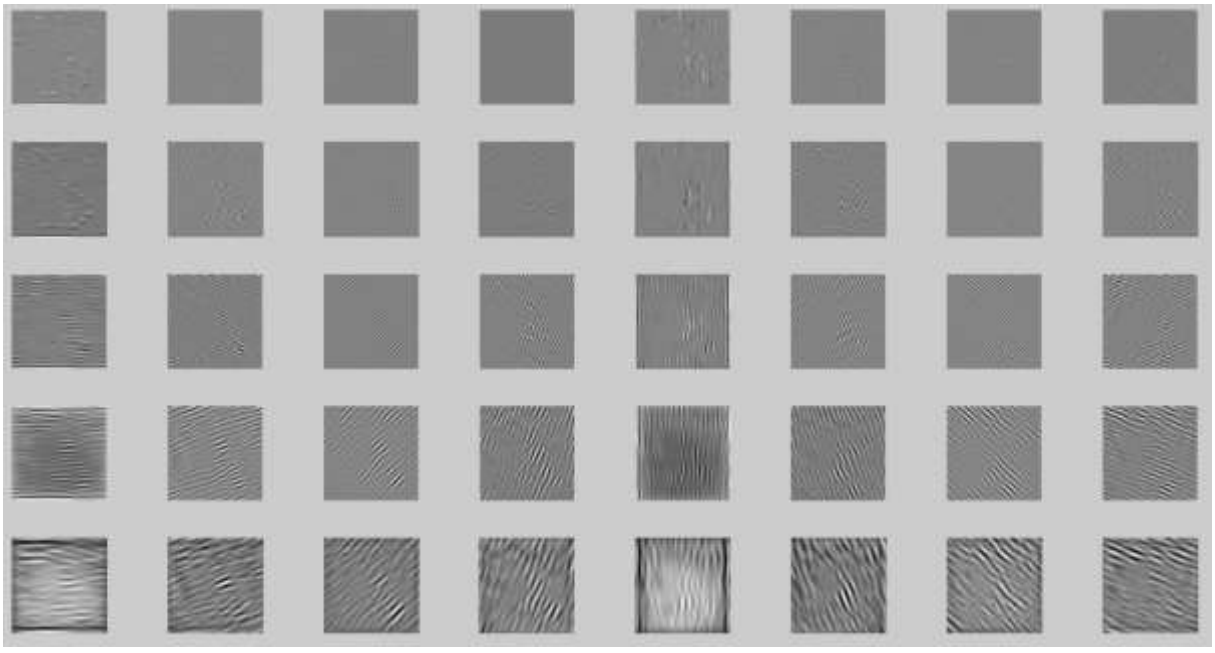

Figure 17:Parties réelles des régions d'intérêt filtrées

A présent, nous allons présenter sur la figure 18, la magnitude de la réponse des régions d'intérêt filtrées.

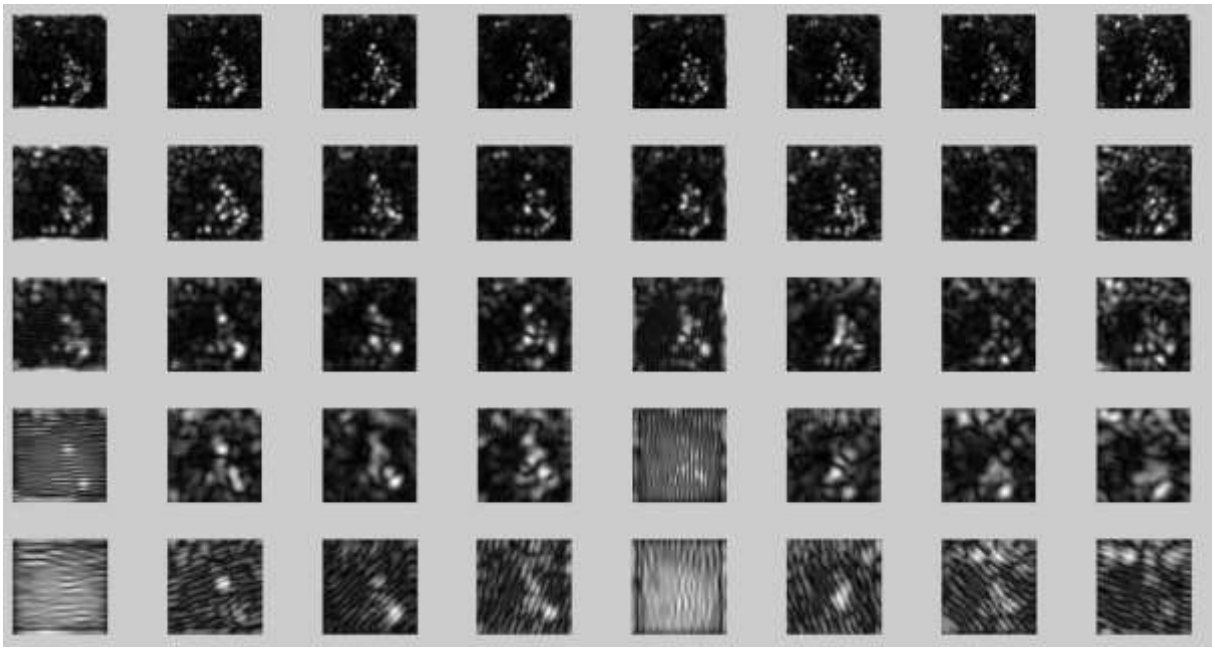

Figure 18:Magnitude de la réponse du filtre de Gabor après avoir effectué la convolution de la tumeur maligne avec une banque de filtres de Gabor.

A partir de ces résultats, un calcul de statistiques locales est effectué pour toutes les sorties du filtre et le vecteur de paramètres texturaux est obtenu par combinaison de ces statistiques comme mentionné dans [72].

3.4. Processus d'apprentissage automatique

Cette phase du pipeline est la phase d'apprentissage proprement dite. En effet, les descripteurs extraits à partir de l'analyse des régions d'intérêt par les lois puissance ainsi que le filtres de Gabor sont introduits dans l'algorithme d'apprentissage automatique choisi ou aux procédures de sélection de descripteurs pour maintenir le sous-ensemble de descripteurs qui contribue le plus à la précision de classification tout en facilitant cette tâche [73].

Par la suite, nous entreprenons un processus itératif pour identifier les modèles appropriés dans les données [35]. En effet, dans le cadre du processus de développement des modèles de classification, il est souhaitable d'essayer divers techniques d'apprentissage automatique pour trouver celle aboutissant à la meilleure valeur prédictive [73].

Une fois l'algorithme termine son optimisation, le modèle d'apprentissage automatique est prêt pour le test. En effet, nous visons à tester la généralisabilité du modèle sur des données non traitées par l'algorithme pendant la phase d'apprentissage. Cette validation est réalisée à l'aide d'un ensemble secondaire de données étiquetées des observations de test qui va passer ensuite à travers les mêmes étapes de prétraitement et d'extraction de descripteurs que dans la phase d'apprentissage [35].

La performance prédictive du modèle est calculée selon la capacité de généralisation mesurée sur cet ensemble, elle est donc d'une importance vitale pour fournir une estimation complète des performances du classifieur [73].

4. Aperçu schématique de l'approche proposée

Nous allons exposer sur la figure 19 un aperçu schématique du pipeline de notre approche proposée basée fusion des lois de puissance : Zipf, Zipf inverse et les filtres de Gabor pour la classification des tumeurs mammaires où nous schématisons toutes les étapes évoquées dans la section précédente.

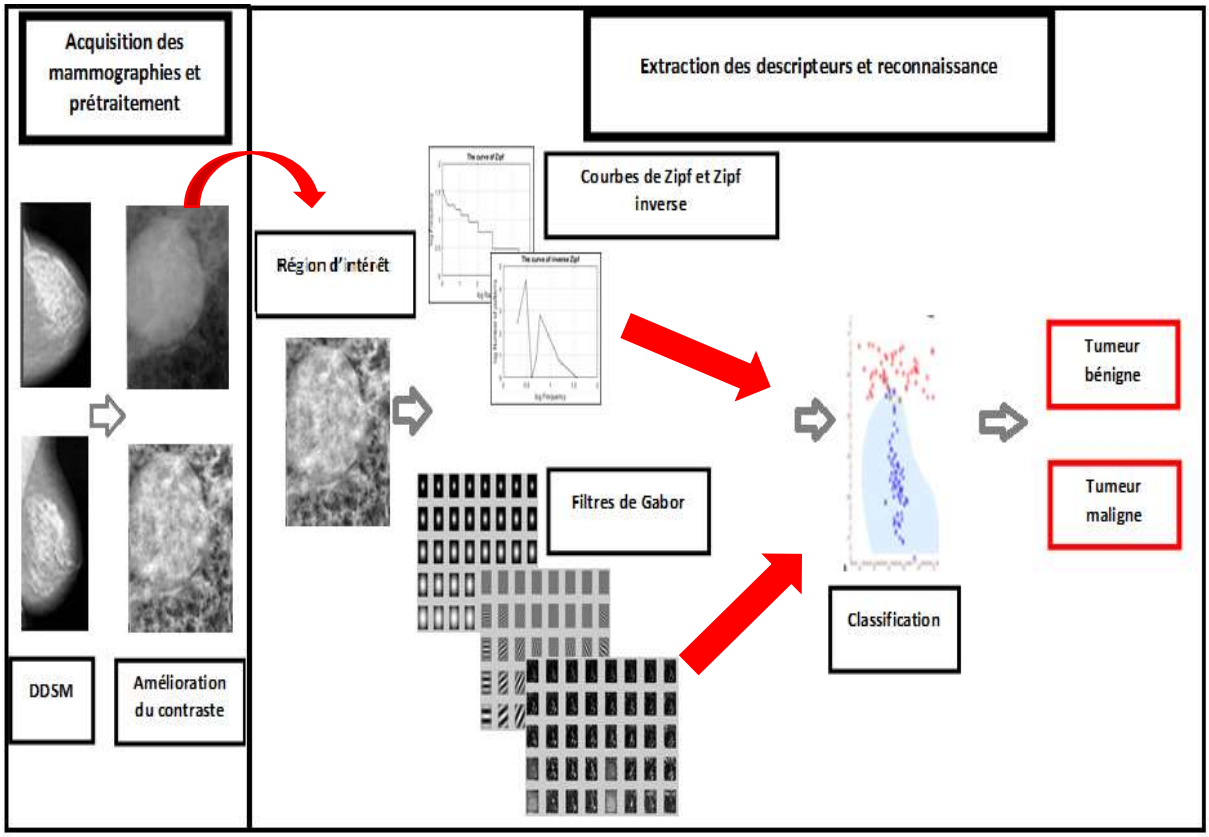

Figure 19:Aperçu schématique du pipeline de notre approche proposée basée fusion des lois de puissance : Zipf, Zipf inverse et les filtres de Gabor pour la classification des tumeurs mammaires.

5. Evaluation de l'approche proposée

5.1. L'environnement de développement

Nous avons implémenté notre approche sur un ordinateur Lenovo 20236 sous le système d'exploitation Windows 10. Processeur Intel Core (TM) i3. Fréquence : 2,5 GHz. Mémoire installée (RAM): 4 Go et la version premium: 64 bits.

5.2. Les outils utilisés

- MATLAB (MATrix LABoratory)

Est un environnement puissant, complet et facile à utiliser destiné au calcul scientifique. En effet, les utilisations de MATLAB incluent les calculs matriciels, le développement et l'exécution d'algorithmes, la création d'interfaces utilisateur (UI) et la visualisation de données. L'environnement de calcul numérique multi-paradigme permet aux développeurs de s'interfacer avec des programmes développés dans différents langages, ce qui permet d'exploiter les atouts uniques de chaque langage à des fins diverses.

Mentionnons que les ingénieurs et les scientifiques utilise MATLAB à travers divers domaines comme le traitement d'image et le traitement du signal, les systèmes de contrôle dédiés à l'industrie, la conception de réseaux intelligents, la robotique ainsi que la finance informatique.

- Python

Python consiste en un langage de programmation interprété, orienté objet et de haut niveau doté d'une sémantique dynamique. Ses structures de données intégrées et de haut niveau, combinées au typage dynamique ainsi qu'à la liaison dynamique, l'ont rendu hyper attractif durant le développement rapide d'applications, ainsi que pour une utilisation comme langage de script pour connecter des composants entre eux.

Citons qu'Il est évident d'implémenter sous Python en accédant à des bibliothèques de haute qualité. En effet, il est omniprésent dans le domaine de l'intelligence artificielle et de l'apprentissage automatique. Pandas, image-Net, pascal et mlxtend prennent en charge Python et sont disponibles pour de plusieurs systèmes d'exploitation.

Les bibliothèques les plus utilisées dans l'intelligence artificielle sont implémentées par Python ; citons opencv et Scikit-learn et bien beaucoup d'autres.

5.3. Evaluation des performances

Selon les données collectées, il est difficile de comparer de manière exhaustive les méthodes entre elles en raison de plusieurs facteurs. Certains de ces facteurs sont [44]:

- Les bases de données utilisées pour l'évaluation.
- Les échantillons d'images sélectionnés pour l'évaluation.
- Le nombre d'échantillons utilisés.
- L'approche d'évaluation (méthodologie de validation, formation et ensemble de tests) utilisée.
- Le réglage des paramètres impliqués dans différentes méthodes varie d'une méthode à l'autre, ajoutant ainsi un autre obstacle à une comparaison équitable entre les différentes méthodes.

Étant donné que la plupart des algorithmes de classification s'exécutent sur de petits ensembles de données, ils ne seront pas adaptés pour identifier et classer de grands ensembles de données dans les hôpitaux [74]. De plus, la précision ne sera pas comparable. En effet, les performances de généralisation dépendent fortement sur la taille de l'ensemble de d'apprentissage par rapport à la complexité du modèle de représentation et d'apprentissage [75].

Voir plus, étant donné que l'ensemble de données soit relativement petit, les métriques d'apprentissage automatique standard seules ne seront pas en mesure de fournir une confiance accrue dans l'évaluation des résultats obtenus [62]. Donc, la taille (et la qualité) de l'ensemble d'apprentissage est un facteur d'une importance vitale. Essentiellement, collecter plus de données et les rendre disponibles pour l'apprentissage aidera à réduire l'écart entre l'erreur d'apprentissage et de la généralisation. En effet, nous visons à éviter l'overfitting [33], c'est-à-dire lorsque les paramètres q produisent une fonction f qui est très précise pour les données d'apprentissage mais qui ne se généralise pas bien aux données de test.

Dans le domaine médical, la situation est plus compliquée [73], en particulier pour les évaluation de la progression de la maladie ou de toute autre tâche pronostique. Dans la plupart des cas, les ensembles de données sont composés d'un nombre limité de dossiers de patients avec des informations pronostiques complètes. L'une des approches les plus populaires pour faire face à ce problème est la validation croisée. Cette technique statistique divise au hasard les échantillons de données en n plis disjonctifs, où n est défini au préalable. Ensuite, dans les étapes de procédure à n itérations, n_ 1 plis sont utilisés pour l'entraînement, et le dernier est

utilisé pour les tests. Les performances du classifieur sont la valeur moyenne des performances obtenues à chaque étape de l'itération. Il est important de permettre une proportion égale de classes dans chaque pli. En outre, la variance des performances du classifieur pourrait encore être réduite en répétant la validation croisée avec différents échantillons aléatoires de données.

Les algorithmes d'apprentissage non supervisé et supervisé peuvent être appliqués à différents types de données médicales [76], y compris les données cliniques, histologiques ainsi qu'images médicales. En effet, trouver le bon algorithme d'apprentissage automatique n'est acquis que suite à de divers essais [31]. En effet, la sélection des algorithmes est basée sur le type et taille des données et les sorties requises à partir des données d'entrée.

En tirant profit de tout ce qui a été cité dans les paragraphes précédents, nous avons travaillé avec deux ensembles de données. Le premier ensemble de données contient 2000 régions d'intérêts : 1000 régions d'un tissu normal et 1000 autres de tumeurs malignes et bénignes pour le processus de classification : tissu normal ou une tumeur où nous démontrerons la puissance de notre vecteur descripteur à distinguer un tissue pathologique d'un tissue parenchyma normal. En effet, nous avons commencé par la distinction entre les régions d'intéret portants un tissue normal et celles portant une tumeur durant le premier stade de nos experimentations.

Quant au deuxième ensemble de donnée, nous prendrons 1000 régions d'intérêt à partir du premier ensemble de données: 500 régions d'intérêts portants des tumeurs bénignes et 500 autres régions d'intérêts portant des tumeurs malignes; ceci pour le processus de classification en deux classes: tumeurs malignes ou tumeurs bénignes; désignant le stade principal de notre approche. Rappelons que le processus d'acquisition des régions d'intérêt a été évoqué dans la section 3.2.
Les données de test ainsi que d'apprentissage sont divisées selon le rapport de division standard 80:20 [77] et une métrique d'évaluation est ensuite utilisée pour évaluer l'exactitude du modèle sur la base des véritables étiquettes des données de test. Les paramètres d'évaluation populaires pour la classification sont basés sur la matrice de confusion comme suit [73] :

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Où nous calculons à partir de cette matrice, les mesures suivantes :

$$Acc = \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad (VI.7)$$

$$Sens = \frac{TP}{(TP+FN)} \quad (VI.8)$$

$$Spec = \frac{TN}{(TN+FP)} \quad (VI.9)$$

Où « Accuracy » est la mesure la plus utilisée pour l'évaluation d'un modèle [73], désignant la proportion du nombre total de prédictions qui étaient correctes: TP et TN présentent le nombre d'instances positives et négatives correctement classées alors que FN et FP représentent le nombre d'erreurs de classification des valeurs positives et négatives.

Nous allons présenter sur le tableau suivant la comparaison des résultats obtenus avec des approches de l'état de l'art pour la classification des régions d'intérêt comme tissue normal ou tumeur:

Tableau 2:Comparaison de l'approche proposée vis-à-vis des approches de l'état de l'art pour la classification basée tissu normal ou tumeur

Référence	**Technique d'apprentissage Automatique**	**Descripteurs**	**Evaluation des résultats**	**Base des mammog-raphies utilisée**
[78]	KNN	Curvelet	Acc = 91.27%	Mini-MIAS 252 DDSM 11553
[79]	SVM	Gabor+PCA	Acc = 84%	DDSM Nbr NM
[80]	SVM	HOG+DSIFT+LCP	Acc = 84%	DDSM 600
[81]	Multi-layer Perceptron (MLP)	Zernike	Acc = 82.95%	DDSM
[82]	SVM	MLPK	Acc = 86%	DDSM 584

Tableau 2 (suite)

Référence	Technique d'apprentissage Automatique	Descripteurs	Evaluation des résultats	Base des mammographies utilisée
[83]	BPNN	WGLCM	Acc = 88.33%	DDSM
[57]	Discrimination Potentiality	DP-HOT	Acc = 82.56%	DDSM 2576
[57]	Discrimination Potentiality	DP-PB-DCT	Acc = 84.84%	DDSM 2576
Suggestion	SVM	Zipf et Zipf inverse	Acc =73.1%	DDSM 2000
	SVM	Gabor	Acc =69.95%	DDSM 2000
Approche proposée	SVM	Zipf et Zipf inverse+Gabor	Acc = 85%	DDSM 2000
	Random forest		Acc = 88%	

Nous allons présenter sur le tableau suivant la comparaison des résultats obtenus avec des approches de l'état de l'art, mais à présent pour la classification des 1000 régions d'intérêt comme tumeurs bénignes ou tumeurs malignes:

Tableau 3:Comparaison de l'approche proposée vis-à-vis des approches de l'état de l'art pour la classification basée tumeurs malignes ou bénignes

Référence	**Technique d'apprentissage Automatique**	**Descripteurs**	**Evaluation des résultats**	**Base des mammog-raphies utilisée**
[84]	KNN	Transformée de Hough	Acc = 65%	DDSM 615
[85]	ANN	SGLD	Acc = 71%	DDSM 377
[86]	KNN	GLCM	Acc = 79%	DDSM 144
[87]	SVM	Wavelet	Acc = 73%	DDSM 480
[88]	SVM	texture + géométrie	Acc = 94%	DDSM 826
[79]	SVM	Gabor+PCA	Acc = 78%	DDSM Nbr NM
[80]	SVM	HOG+DSIFT+LCP	Acc = 78%	DDSM 600
[81]	(DT, SVM, KNN)	Descripteurs de contours de la tumeur	Acc = 72%	DDSM
[83]	BPNN	WGLCM	Acc = 55.09%	DDSM
[57]	Discrimination Potentiality	DP-PB-DCT	Acc = 72.43%	DDSM
[89]	LDA	Descripteurs statistiques	Acc = 78.57%	DDSM 168
[89]	QDA	Descripteurs statistiques	Acc = 76.19%	

Tableau VI.2 (suite)

Référence	Technique d'apprentissage Automatique	Descripteurs	Evaluation des résultats	Base des mammographies utilisée
[90]	SVM	Gabor wavelet	Acc = 78.26%	DDSM 114
Suggestion	SVM	Zipf et Zipf inverse	Acc =65.3%	DDSM 1000
	SVM	Gabor	Acc =61.87%	DDSM 1000
Approche proposée	SVM	Zipf et Zipf inverse+Gabor	Acc = 80%	DDSM 1000

Nous avons réalisé diverses itérations répétitives durant la phase d'apprentissage et de test avec différents hyperparamètres (liés à la complexité du modèle : pour le RBF kernel des SVM avec des valeurs variées de C=1000, γ=0,01 par exemple) ; où l'ultime but était d'aboutir à la complexité optimale pour un algorithme et un ensemble de données donnés.
Pour la 1ère phase de la classification basée tissu normal ou tumeur où les résultats obtenus étaient évoqués sur le tableau 2, en effet, les précisions de classification étaient de l'ordre de 85% et 88% pour les SVM et le random forest respectivement pour un ensemble de donnée de 2000 régions d'intérêt.

Pour la 2ème phase de classification basée tumeur maligne ou bénigne où les résultats obtenus étaient évoqués sur le tableau 3, en effet, nous avons obtenu une précision de classification de l'ordre de 80% en prenant les 1000 régions d'intérêt portants des tumeurs malignes et bénignes à partir des 2000 régions d'intérêt.

Depuis les tableaux comparatifs nous énumérons les attestations suivantes :

- Lorsque les ensembles de données sont relativement petits, le classifieur a tendance à apprendre et non pas à généraliser (ici on parle du overfitting). Donc il est très important que l'approche de l'apprentissage automatique soit basée sur un large ensemble de données (notre cas : 2000 pendant la 1ère phase et 1000 pendant la seconde) ; au lieu de

la majorité des approches de l'état de l'art à utiliser uniquement un petit ensemble de données, surtout que les hôpitaux à qui ces approches sont dédiées obtiennent de très large ensembles de données par jour.

- Divers travaux adoptent la fusion des descripteurs pour la caractérisation des tumeurs où notre fusion des lois de Zipf et Zipf inverse ainsi que les filtres de Gabor a surpassait la majorité des approches de l'état de l'art exposées sur les tableaux à l'exception de l'approche présentée dans [88] où la fusion était basée sur des approches texturaux ainsi que d'autres géométriques d'où l'idée de rajouter ultérieurement les descripteurs géométriques dans notre vecteur descripteur pour atteindre les meilleurs performances.
- Il est relativement difficile de classifier les régions d'intérêt en tumeur bénignes ou malignes par rapport à la classification en tissue normal ou tumeur en raison du manque des propriétés distinctives comme il a été affirmé dans [91].
- Il y a des risques associés à l'application des techniques d'apprentissage comme une «boîte noire» pour effectuer le diagnostic et l'évaluation des risques. Un système d'apprentissage flexible dans un espace de descripteurs de grande dimension doit être basé sur l'essai de divers classifieurs avec différents paramètres pour en choisir le meilleur. En effet, en raison de ces problèmes, il ne suffit pas de savoir qu'une approche d'apprentissage a d'excellentes performances sur un ensemble de données donné mais nous devrions viser à comprendre quelles caractéristiques motivent les décisions et quels sont les pièges correspondants.

6. Conclusion

Nous avons présenté tout au long de ce chapitre l'approche proposée ainsi que la validation des résultats obtenus. En effet, la comparaison des résultats de ce travail avec les travaux présentés dans l'état de l'art n'était pas une tâche simple car ils emploient des méthodologies, bases de données d'images et nombre d'échantillons pour des tests, différents. Cependant, nous avons pu démontrer la puissance de la fusion des lois de puissance : Zipf et Zipf inverse avec les filtres de Gabor pour une caractérisation pertinente des images mammaires.

Conclusion générale

À ce jour, les applications de l'apprentissage automatique à la santé numérique ont été limitées à la recherche et aux milieux universitaires. Néanmoins, quelques avancées récentes vers ces objectifs ont été atteintes par l'industrie [35]. En effet, Bien que les performances des systèmes CAD soient modérées, ils peuvent détecter des lésions de différentes caractéristiques que celles détectées par des radiologues voir plus, il a été prouvé que la sensibilité globale augmente lorsque le radiologue lit les mammographies côte à côte avec le CAD où les études ont montré que la précision des radiologues était significativement améliorée lors de la lecture avec CAD [75].

Dans les images mammaires, nous distinguons des lésions obscurcies possédant majoritairement des apparences semblables, ceci est un inconvénient par rapport à la suffisance discriminatoire des descripteurs géométriques pour la distinction entre les masses bénignes et malignes. Dans ce sens, un bon remède à la caractérisation optimale des lésions mammaire serait la texture. En effet, l'analyse de la texture est cruciale vis-à-vis la vision humaine, cependant, les descripteurs texturaux sont difficilement perçues et quantifiés par la vision humaine.

Nous avons caractérisé les propriétés texturales des motifs à diverses fréquences et orientations pour une meilleure séparabilité entre les différentes fonctionnalités extraites. En effet, la sortie d'énergie du filtre de Gabor de chaque réponse en amplitude était combinée avec les descripteurs extraits à partir des lois de Zipf et de Zipf inverse.

Les approches d'apprentissage automatique semblent prendre le dessus et réussissent de plus en plus sur le domaine de diagnostic basé sur l'image et assisté par ordinateur [75]. Ceci dit, de nombreux défis pratiques et scientifiques doivent encore être relevés pour débloquer leur plein potentiel. Citons : un moyen de former des modèles solides sur peu de données, l'amélioration de l'accès aux données, comment utiliser au mieux la structure de l'image et les propriétés spécifiques de l'imagerie médicale au cours de la conception des modèles, comment interpréter les résultats et comment appliquer ces résultats dans la pratique clinique.

Ainsi que d'autres obstacles à surmonter, nous soulignons les points suivants :

- La poursuite des recherches dans le domaine de l'IA pour la santé numérique dans le but de construire des plates-formes commerciales liées à la santé. En effet, l'aide au diagnostic médical assisté par ordinateur (CAD) est d'une grande utilité pour le soutient des contrôles médicaux préventifs en mammographie ainsi que pour d'autres domaine comme la neurologie ou le cardio-vasculaire.

- Faciliter l'accès aux données médicales qui n'est pas évident en raison de problèmes de confidentialité, ce qui complique le partage des données collectées entre les groupes de chercheurs pour comparer les méthodologies contribuées et les résultats aboutis.
- Des études ont indiqué que les informations cliniquement significatives ne sont pas seulement concentrées sur la lésion, mais se répartissent également sur l'intégralité de la zone mammaire sur la mammographie; en effet, il est difficile à identifier de manière adaptative la taille optimale des régions d'intérêt pour couvrir les lésions de taille et de forme variables [75].
- Les approches d'apprentissage automatique ont des limites sur le point que le développeur humain peut ne pas être en mesure de traduire les structures complexes de la maladie en un nombre de descripteurs.

Références

[1] Christine , G . Annick , M . thése : Analyse d'images: Filtrage et segmentation, 2017

[2] Ramla , I . Asmae Mama , Z , Mémoire de Master : Detection des lesions de l'abdomen et reconstruction tridimensionnelle, 2016.

[3] Mohammed Habib , B thèse : Développement de méthodes d'extraction de contours sur des images à niveaux de gris , 2017

[4] Laurent , G ; thèse de l'Université de Cergy-Pontoise : Modèles Multi-Échelles pour la Segmentation d'Images ; Décembre 2003

[5] Meriem , H , Indexation et segmentation d'images basées loi de Zipf et Zipf inverse, mémoire de doctorat ism 2015

[6] Belkis , B . Hayet , D Mémoire de Master : Analyse de la texture des images mammaires par une fusion des lois de Zipf et des SFTA pour la classification des tumeurs mammaires via l'analyse en composantes principales ,mémoire de master 2018.

[7] Brian , L. DeCost , Elizabeth , A . Holm. A computer vision approach for automated analysis and classification of microstructural image data. Computational Materials Science, (2015).

[8] Jianguo Z , Tieniu T ,Brief review of invariant texture analysis methods 2001

[9] Hai, J.; Tan, H.; Chen, J.; Wu, M.; Qiao, K.; Xu, J., et al. Multi-level features combined end-to-end learning for automated pathological grading of breast cancer on digital mammograms. Computerized Medical Imaging and Graphics (2019)

[10] Chahnez Hadj S , Khadidja B : Filtre de Gabor. Université Abou Bekr Belkaid – Tlemcen. 2014/2015

[11] Chabha ,T . Assia , Z . Mémoire de fin d'étude : Segmentation d'images texturées par filtrage de Gabor : applications aux images médicales. 2011.2012

[12] Jacques Philémon , M : Apports de la texture multibande dans la classification orientée-objets d'images multisources.Octobre 2016.

[13] Maroua , M . Mohamed , M . Petra , Gomez-K . Pierre ,H . Mohamed Ali , M . et al.. Étude comparative de trois ensembles de descripteurs de texture pour la segmentation de documents anciens. CIFED 2014 - Actes du treizième Colloque International Francophone sur l'Écrit et le Document, Mar 2014, Nancy, France. pp.41-56, 2014.

[14] Haralick , R M.. Shanmugam , K. Dinstein , I. :Textural features for image classification , SMC, vol. 3, no 6, p. 610-621, 1973.

[15] Hedjaz H , Doctorat 3ème Cycle en LMD 2018, Identification de personnes par signature manuscrite.

[16] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid: Local features and kernels for classification of texture and object categories: comprehensive study. International Journal of Computer Vision (Springer). Vol.73, No.2, pp.213-238.

[17] George.K. Zipf, Human Behavior and the Principle of Least Effort. Addison-Wesley, New York, 1949.

[18] Caron Y., Makris P., Vincent N.,(2007), Use of power law models in detecting region of interest, Pattern Recognition, Elsevier 40 (9): 2521– 2529.

[19] Lavalette D., (1996), On a Zipf's law extension to impact factors, INSERM U350 Institut curie Recherche, Bt 112, Centre Universitaire 91405, 1996, Orsay, France.

[20] Benguigui L , Blumenfeld L.E., Beyond the power law – a new approach to analyze city size distribution (2007). Computers environ urban system 31(6):648–666.

[21] Lei S., Zhimin G., Lin Wand Y.S., An applicative study of Zipf's law on web cache (2006),. International Journal of information and technology 12(24).

[22] Meriem H, Hayat Farida M., Lakhdar L ,The power laws: Zipf and inverse Zipf for automated segmentation and classification of masses within mammograms. Evolving systems. Springer, Berlin, 2014.

[23] Caron Y., Makris P., Vincent N., (2002), A method for detecting artificial objects in natural environments, RFAI team publication international conference on pattern recognition ICPR 2002, Quebec, pp. 600–603.

[24] Vincent N., Makris P., Brodier M., (2000), Compressed image quality and Zipf's law. In: International conference on signal processing (ICSP–IFICIAPR WCC2000), Beijing (China) 1077–1084.

[25] Lakhdar L, Hayat Farida M , A novel Technique of Steganalysis in Uncompressed Image through Zipf's Law (2012), In International Journal of Computer Applications, 40 (6): 0975-8887.

[26]Caron Y, Contribution de la loi de Zipf à l'analyse d'images, thèse doctorat, L'UNIVERSITE DE Tour 2004

[27] Sandra ,V. Walter Hugo, L P . Andrea , M . Introduction to machine Learning . Machine Learning Elsevier , 2020 pp.1-20

[28] Sandra, V. Walter Hugo, L P. Andrea , M . Machine learning techniques . 2020 pp.91-202

[29] Chloé-Agathe , A . Introduction au Machine Learning (Dunod, 2018).

[30] Jair ,C . Farid ,Garcia-L . Lisbeth , Rodríguez-M . Asdrubal , L. A Comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 2020.

[31] Rohit ,T . Surekha , B . Application of Machine Learning Algorithms for Classification and Security of Diagnostic Images . Machine Learning in Bio-Signal Analysis and Diagnostic Imaging. 2019, pp. 273-292

[32] V. N. Vapnik and S. Kotz, Estimation of dependences based on empirical data, vol. 40. Springer Verlag New York, 1982.

[33] Ashnil , K . Lei , B. Jinman , K . David Dagan , F . Machine learning in medical imaging. Biomedical Information Technology 2020, pp 167-196

[34] Rachid , M . Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers Mémoire à l'école High-Tech 2019.

[35] Nicholas , C. Zhao , R . Adria , Mallol-R . Bjorn , S . Machine learning in digital health, recent trends, and ongoing challenges. Artificial Intelligence in Precision Health. 2020, pp 121-148

[36] Abir , S . When machine learning meets medical world: Current status and future challenges. Computer Science Review 2020.

[37] Andy M.Y, T. Alcides , A . Nicole E , C . Mehala , S . Danielle S, C . Margarita , S . Yena , L . Rodrigo , M . Roger S, McIntyre. Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. Artificial Intelligence In Medicine 2019.

[38] Gobert_, L. Hiroshi , F Deep Learning in Medical Image Analysis. Editors. Advances in Experimental Medicine and Biology. 2020.

[39] Sujatha , K. Durgadevi, G. Senthil Kumar, K. Karthikeyan , V. Ponmagal , R.S. Rajeswari Hari Bhavani, , N.P.G. Srividhya ,V. and Su-Qun Cao. Screening and early identification of microcalcifications in breast using texture-based ANFIS classification.. Wearable and Implantable Medical Devices. 2020 pp 115-140.

[40] Mohamed, H . Ibtissam , A. Ali , I . Juan M , C de G , JoséLuis Fernández , A . Reviewing ensemble classification methods in breast cancer. Comput Methods Programs Biomed 2019 pp 89-112

[41] Mohammed , N. Amine, A . Bouchra ,N . Hanaa , H . A highly efficient system for Mammographic Image Classification Using NSVC Algorithm. Procedia Computer Science 2019 pp 135-144

[42] Muhammad , K .Kaleem , R M , Sohail , J and Junaid , C . Application of machine learning and image processing for detection of breast cancer. Innovation in Health Informatics. 2020 pp 145-162.

[43] Simara ,Vieira da R . Geraldo Braz , J . Aristófanes , C . Anselmo , C P . Marcelo , G . Texture Analysis of Masses Malignant in Mammograms Images Using a Combined Approach of Diversity Index and Local Binary Patterns Distribution, Expert Systems With Applications 2016 pp 7 -19

[44] Nisreen , I.R. Yassin . Shaimaa , O . Enas M.F. El H , Hemat , A. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities . Comput Methods Programs Biomed 2018 pp 25-45

[45] Jalalian, A. Mashohor, S. B .T. Rozi Mahmudb, H. Iqbal. M. Saripan, B. Ramli, A. B. Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clinical Imaging, 37. pp. 420–426.

[46] Jiang. Y. Handbook of Medical Imaging . Classification of Breast Lesions from Mammograms. 2000 pp. 341-357.

[47] Mugahed , A . Al-antari Ph , D . , Tae-Seong Kim , Ph D , Evaluation of Deep Learning Detection and Classification towards Computer-aided Diagnosis of Breast Lesions in Digital X-ray Mammograms, Computer Methods and Programs in Biomedicine 2020

[48] Lobbes, M. Smidt , M . Keymeulen , K . Girometti , R. . Zuiani , C. Beets-Tan , R. Wildberger , J. Boetes , C . Malignant lesions on mammography: accuracy of two different computer-aided detection systems. 2013 pp 283-288 .

[49] Leon , S. Libby , B . Honeyman-Buck J., Marshall , J. Comparison of two commercial CAD systems for digital mammography. Journal of Digital Imaging, 2009 pp 421–423

[50] Bartosz , S . Stanislaw , O . Jaroslaw , K . Michal , K . Iwona Lugowska , Piotr , R. Walid , B . Novel methods of image description and en-semble of classiÞers in application to mammogram analysis, 2017

[51] Dhahbi , S. Barhoumi ,W. Zagrouba , E. Breast cancer diagnosis in digitized mammograms using curvelet moments. Computers in Biology and Medicine, 2015. pp79-90.

[52] Vijayarajeswari, R . Parthasarathy, P. Vivekanandan, S. Alavudeen Basha , A . Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measurement 2019, pp 800-805.

[53] Debendra , M. Ratnakar , D . Banshidhar , M . Automated breast cancer detection in digital mammograms:A moth flame optimization based ELM approach. Biomedical Signal Processing and Control 2020

[54] Sk Md , O. Ahmed , S . Teresa , G . Lu´ıs , R. RMID: A Novel and Efficient Image Descriptor for Mammogram Mass Classification. 2018, pp. 229–240

[55] Omid , Rahmani S . Javad , H . Evaluation of a new ensemble learning framework for mass classification in mammograms, Clinical Breast Cancer 2017

[56] Muhmmad Irfan , S . Jian Ping , L . Javeria , N . Rashid , Iqra A Comprehensive Review on Multi-Organs Tumor Detection based on Machine Learning, Pattern Recognition Letters 2019 pp30-37

[57] Aditya ,A S . Deepti T . Kapil A . Density-wise two stage mammogram classification using texture exploiting descriptors. Expert Systems With Applications 2018 pp 71-82.

[58] Heminder , K . Jitendra , V. K , Shruti , T. A genetic algorithm-based metaheuristic approach to customize a computeraided classification system for enhanced screen film mammograms. U-Healthcare Monitoring Systems 2019 pp 217-259

[59] Gupta , S. Zhang , D. Sampat , M.P. Markey , M.K. Combining texture features from the MLO and CC views for mammographic CADx, in: Medical Imaging 2006: Image Processing, 2006.

[60] Sapate S . Sanjay ,T . Abhishek , M . Nilesh , S. Subhash ,D . Meenakshi ,T. . Breast cancer diagnosis using abnormalities on ipsilateral views of digital mammograms. Biocybern Biomed 2019 pp290-305.

[61] Amira , J , Abir , B . Walid , B . Multi-view information fusion in mammograms: A comprehensive overview. Information Fusion 2019 pp 308-321

[62] Suhas , G S . Abhishek , M . Sanjay , N T . Nilesh , S . Subhash . D , Meenakshi , T ,Radiomics Based Detection and Characterization of Suspicious Lesions on Full Field Digital Mammograms, Computer Methods and Programs in Biomedicine 2018 pp 1-20

[63] Nagarajan , V. Britto ,EC . Veeraputhiran , SM, Feature extraction based on empirical mode decomposition for automatic mass classification of mammogram images, Medicine in Novel Technology and Devices, 2019

[64] Indrajeet , K . Jitendra V . Harvendra S. Bhadauria, Manoj K. Panda and Kriti. Classification of Breast Density Patterns Using PNN, NFC, and SVM Classifiers. Chapter. Soft Computing Based Medical Image Analysis. 2018. pp 223-243

[65] Heath, K B M. Kopans, D. Moore, R. Kegelmeyer, W P. The digital database for screening mammography, Medical Physics, 2001 pp. 212-218 .

[66] Fernando Soares , S O. Antonio Oseasde Carvalho, F . Aristófanes , C S, . Anselmo , C P . Marcelo , G. Classification of breast regions as mass And non-mass based on digital mammograms using taxonomic indexes and SVM. Computers in Biology and Medicine 2015 pp 42-53

[67] Şaban , Ö . Bayram ,A. Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM and SFTA. Procedia Computer Science 2018 pp 40-46

[68] Omar Sultan Al-K. a gabor filter texture analysis approach for histopathological brain tumour subtype discrimination.2017.

[69] Belkacem , A . Lyes , M. Analyse de texture par les filtres de Gabor et Laws. Mémoire de Master en Electronique. 2014

[70] Olivier , R. Méthodes d'analyse de texture pour la cartographie d'occupations du sol par télédétection très haute résolution : application à la fôret, la vigne et les parcs ostréicoles. Traitement du signal et de l'image [eess.SP]. Université de Bordeaux, 2014.

[71] Salem,W . North C . image classificaton using gabor filters and machine learning. In Partial Fulfillment of the Requirements for the Degree of master of science 2009

[72] Haghighat, M. Zonouz, S. Abdel-Mottaleb M. CloudID: Trustworthy cloud-based and cross-enterprise biometric identification". Expert Systems with Applications, 2015 pp. 7905-7916.
[73] Filipovic. Machine learning approach for breast cancer prognosis prediction. Computational Modeling in Bioengineering and Bioinformatics

[74] Hua , L . Shasha , Z . Deng-ao, L . Jumin , Z . Yanyun , Ma. Benign and malignant classification of mammogram images basedon deep learning. Biomedical Signal Processing and Control 2019.pp 347-354

[75] Marleen , de B . Machine learning approaches in medical image analysis: from detection to diagnosis, Medical Image Analysis 2016.

[76] Vasileios , P .Themis , E . Dimitrios , F. Medical Data Sharing, Harmonization and Analytics 2020

[77] Tilottama G. Machine learning behind classification tasks in various engineering and science domains. Cognitive Informatics, Computer Modelling, and Cognitive Science, 2020 pp 339-356

[78] Dhahbi, S. Barhoumi, W. Zagrouba, E. Breast cancer diagnosis in digitizedmammograms using curvelet moments, Comput. Biol. Med. 2015pp 79–90.

[79] Buciu, I. , & Gacsadi, A. Directional features for automatic tumor classification of mammogram images. Biomedical Signal Processing and Control, 2011 pp 370–378.

[80] Ergin, S. , & Kilinc, O. A new feature extraction framework based on wavelets for breast cancer diagnosis. Computers in Biology and Medicine, 2015 pp171–182.

[81] Tahmasbi, A. , Saki, F. Shokouhi, S. B. Classification of benign and malignant masses based on Zernike moments. Computers in Biology and Medicine,2011 pp 726–735.

[82] Nanni, L. , Brahnam, S. Lumini, A. A very high performing system to discriminate tissues in mammograms as benign and malignant. Expert Systems with Applications,2012 pp 1968–1971.

[83] Beura, S . Majhi, B. Dash, R. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing, 2015 pp 1–14.

[84] Karssemeijer, N. Automated classification of parenchymal patterns in mammograms, Phys. Med. Biol. 1998 pp 365–389.

[85] Bovis, K. Singh, S. Classification of Mammographic Breast Density using a Combined Classifier Paradigm, In medical image understanding and analysis (MIUA) conference, Portsmouth(C), 2002 pp. 1–4.

[86] Mustra, M G . Delac, K. Breast density classification using multiple features selection, AUTOMATIK, J. Control Meas., Electron. Comput. Commun. 2012 pp 362–372.

[87] Kanisha,B . Lokesh, S . Kumar, P M. Parthasarathy, P. Chandra Babu, G. Speech recognition with improved support vector machine using dual classifiers and cross fitness validation, Pers. Ubiquitous Comput. 2018 pp 1–9.

[88] Liu , X . Tang, J ,Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method , 2014 pp. 910-920.

[89] M. A. Al-antari . Mohammed , A Al-masni . Sung, U , Park. JunHyeok , P. Mohamed , K Metwally . Yasser, M Kadah . Seung-Moo, H . Tae-Seong, Kim . An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network, J. Med. Biol 2017 pp 443–456

[90] Ioan , B . Alexandru, G .“Directional features for automatic tumor classification of mammogram images,” Biomedical Signal Processing and Control ,2011 pp. 370-378.

[91] Mudigonda, N R. . Rangayyan ,R. M. Desautels, J. E. L. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. IEEE Transactions on Medical Imaging, 2001 pp 1215–1227 .