République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaia
Faculté des Sciences Exactes
Département d'Informatique

MÉMOIRE DE MASTER

EN VUE DE L'OBTENTION DU DIPLÔME DE MASTER

Domaine : Mathématiques et Informatique **Filière :** Informatique

Spécialité : Intelligence Artificielle

Présenté par

M. HOUARI Yazid et M. IDIR Adel

Thème

Méthode combinée d'apprentissage pour la classification raffinée du cancer du sein

Soutenu le 10 juillet 2024 devant le jury composé de :

M. ACHROUFENE Achour	**MCA**	**Univ. de Béjaia**	**Président**
M. SLIMANI Hachem	**Professeur**	**Univ. de Béjaia**	**Encadrant**
Mme CHERIFI Feriel	**MCB**	**Univ. de Béjaia**	**Examinatrice**
M. ATTOUMI Hocine	**Doctorant**	**Univ. de Béjaia**	**Examinateur**

Année Universitaire : 2023/2024

⁂ *Remerciements* ⁂

Nous tenons à exprimer notre gratitude à toutes les personnes qui nous ont soutenus et guidés tout au long de ce travail. Nous remercions tout particulièrement notre encadrant, M. SLIMANI Hachem, pour ses précieux conseils et son accompagnement constant.

Nous sommes également reconnaissants envers notre famille et nos amis pour leur soutien inébranlable et leur encouragement tout au long de cette période.

Enfin, nous remercions les membres du jury, M. ACHROUFENE Achour, Mme. CHERIFI Feriel et M. ATTOUMI Hocine, pour avoir pris le temps d'évaluer ce travail.

Merci à tous.

⁕ *Dédicaces* ⁕

Je dédie ce mémoire à mon père, Idriss, dont la sagesse et le soutien m'ont toujours inspiré à persévérer.

À ma mère, pour son amour inconditionnel et son encouragement constant, sans lesquels ce travail n'aurait pas été possible.

Un remerciement tout particulier à ma sœur, Dida, docteur en médecine, qui nous a aidés à écrire le premier chapitre de ce mémoire. Son expertise et son dévouement ont été d'une aide inestimable.

Merci à vous tous pour votre soutien inébranlable et votre inspiration.

M. ***HOUARI Yazid***

Table des matières

Table des figures

Liste des tableaux

Notations et symboles

A	$ANOVA$	Analyse Of Variance
	ANN	Artificial Neural Network
C	CNN	Convolutional Neural Network
D	DRL	Deep reinforcement Learning
G	GAN	Generative Adversarial Network
I	IA	Intelligence Artificielle
	IRM	Imagerie par Résonance Magnétique
K	KNN	K-Nearest Neighbors
M	MHZ	Mégahertz
	ML	Machine Learning
O	OMS	Organisation Mondiale de la Santé
P	PCA	Principal Component Analysis
Q	QL	Q-learning
R	RF	Random Forest
	RL	Reinforcement Learning
S	SVM	Support Vector Machine
	SSL	Semi-supervised Learning
T	TDM	Tomodensitométrie
	TEP	Tomographie par Émission de Positrons
	TNM	Tumor Node Metastasis
U	$UNSL$	Unsupervised Learning

Introduction générale

Le cancer du sein est une maladie caractérisée par la croissance incontrôlée de cellules anormales dans les tissus du sein, formant des tumeurs malignes qui peuvent envahir les tissus environnants et se propager à d'autres parties du corps. En revanche, les tumeurs bénignes du sein, telles que les adénofibromes, sont des masses non cancéreuses qui ne se propagent pas aux tissus voisins ni aux autres parties du corps, bien qu'elles puissent parfois nécessiter un traitement en raison de leur taille ou de leur impact sur la santé.

Le cancer du sein demeure un défi majeur de santé publique à l'échelle mondiale, nécessitant des progrès constants dans les techniques de dépistage précoce et de classification des tumeurs pour améliorer la prise en charge clinique [40].

Dans ce mémoire, nous explorons une méthode combinée d'apprentissage pour la classification raffinée du cancer du sein, en intégrant des avancées significatives en imagerie médicale et en intelligence artificielle (IA). La question qui se pose est : Comment améliorer la précision et l'efficacité de la classification du cancer du sein grâce aux techniques de machine learning et deep learning, tout en surmontant les limitations des méthodes de diagnostic cliniques existantes ? Pour répondre à cette question, nous avons choisi d'explorer la classification des tumeurs du sein en deux catégories (bénigne ou maligne). Ensuite, nous avons effectué une classification des tumeurs malignes en trois stades (précoce, intermédiaire et avancé).

Le premier chapitre établit un cadre conceptuel en abordant les bases du cancer du sein, les avancées en imagerie médicale, et l'émergence de l'IA dans l'analyse d'images médicales. Nous examinons également les méthodes traditionnelles de traitement et les innovations récentes dans le domaine de l'imagerie pour la caractérisation des lésions mammaires.

Le deuxième chapitre se concentre sur les méthodes d'apprentissage automatique pour la détection et la classification du cancer du sein. À travers une revue approfondie de la littérature, nous explorons les approches supervisées, non supervisées, semi-supervisées, par renforcement et combinées, mettant en lumière leurs contributions et leurs limites dans ce domaine crucial de la médecine.

Dans le troisième chapitre, nous introduisons une approche novatrice combinant les réseaux de neurones artificiels (ANN) pour la classification initiale des tumeurs et les méthodes d'apprentissage automatique (RF ou K-means) pour une classification plus détaillée par stades. Cette

méthode hybride promet une évaluation plus précise et nuancée des stades du cancer mammaire, en exploitant la complémentarité des approches basées sur les données tabulaires et les images médicales.

Enfin, le quatrième chapitre évalue cette approche, en analysant les résultats obtenus et en comparant les performances des différentes méthodes évaluées. Nous présentons également les métriques utilisées pour évaluer l'efficacité de notre méthode, concluant par les insights clés et les perspectives futures pour la recherche et la pratique clinique.

Chapitre 1

Généralités sur le cancer du sein et l'analyse d'images

Sommaire

1.1 Introduction

Le cancer du sein demeure un défi majeur de santé publique, constituant l'une des principales causes de décès chez les femmes à l'échelle mondiale. Face à cette réalité, les avancées dans le domaine de l'imagerie médicale et de l'analyse d'images par Intelligence Artificielle (IA) offrent de nouvelles perspectives pour améliorer le dépistage précoce et la prise en charge de cette maladie.

L'intégration de l'intelligence artificielle dans l'analyse d'images représente un domaine de recherche en pleine expansion, offrant des opportunités uniques pour améliorer la précision diagnostique et la personnalisation des traitements. Ce chapitre servira de fondement pour comprendre l'évolution des méthodes de dépistage et de diagnostic du cancer du sein, ainsi que le rôle croissant de l'IA dans ce domaine.

La structure de ce chapitre est organisée de manière à fournir une vue d'ensemble complète, en commençant par les bases du cancer du sein, en passant par les avancées en imagerie médicale, jusqu'à la présentation des principes et des méthodes d'analyse d'images. Cette approche permettra de poser les fondations nécessaires pour une exploration approfondie des sujets abordés dans les chapitres suivants. Plus spécifiquement, dans la Section 1.2, nous commençons par une discussion sur les aspects épidémiologiques du cancer du sein, ses facteurs de risque et les défis associés à son diagnostic et à son traitement. Ensuite, dans la Section 1.3, nous explorerons les avancées technologiques dans le domaine de l'imagerie médicale, en mettant particulièrement l'accent sur les méthodes d'analyse d'images utilisées pour caractériser les lésions mammaires. Dans la Section 1.4, nous passerons en revue la prise en charge thérapeutique d'un cancer du sein ainsi que les différents traitements pratiqués. Dans la Section 1.5, nous plongerons dans le domaine de l'imagerie médicale pour examiner en détail les progrès et les innovations. Nous mettrons particulièrement en lumière les méthodes d'analyse d'images, en nous concentrant sur leur utilisation dans la caractérisation des lésions mammaires.

1.2 Anatomie et épidémiologie

Dans cette partie, nous allons présenter l'anatomie du sein et les tumeurs mammaires. Nous allons également aborder la classification de ces dernières et nous finirons par donner quelques statistiques mondiales.

1.2.1 Anatomie d'un sein normal

Le tissu mammaire comporte des structures épithéliales, constituées de lobules sièges de la sécrétion de lait et des canaux galactophores permettant d'amener le lait au mamelon. Ces structures épithéliales sont entourées de tissu conjonctif, de vaisseaux, de nerfs et de graisse, d'abondance variable selon les femmes, une illustration plus détaillée est présentée dans la Figure 1.1. Lors

des grossesses, une prolifération active des structures épithéliales et du stroma permet d'augmenter le nombre de structures fonctionnelles et, après l'accouchement, en l'absence d'allaitement ou bien après la fin de l'allaitement, une apoptose importante affecte le compartiment épithélial. La graisse remplace progressivement ces structures et, lors de la ménopause, la glande mammaire est en grande partie remplacée par des structures adipeuses et conjonctives. Cependant cette évolution physiologique ne concerne pas toutes les femmes [33].

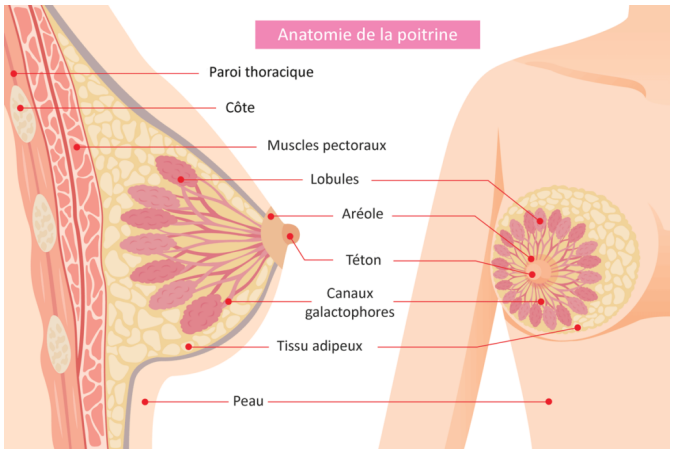


FIGURE 1.1 – Anatomie d'un sein normal [2]

1.2.2 Pathologies tumorales du sein

Une tumeur se caractérise par une augmentation anormale du volume d'un tissu, résultant d'un dysfonctionnement de la croissance cellulaire. Les tumeurs se divisent en deux types : les tumeurs bénignes et malignes, cette dernière étant associée au cancer.

Il existe différents types de tumeurs du sein, spécifiés en fonction du tissu mammaire à partir duquel elles se développent. Certaines sont le résultat de changements bénins, tels que des changements fibrokystiques, des fibroses ou des épaississements du sein. D'autres sont des tumeurs malignes qui commencent généralement dans les cellules canalaires, lobulaires ou d'autres tissus. La formation du cancer du sein peut être déclenchée par plusieurs causes, principalement des facteurs héréditaires et génétiques. En outre, diverses études ont mis en évidence d'autres facteurs majeurs de l'incidence du cancer du sein.

Bien qu'une liste exhaustive de ces facteurs ne soit pas encore entièrement établie, les principaux incluent : les caractéristiques des menstruations (âge précoce de la première règle, âge tardif de la ménopause), la reproduction (nulliparité, âge tardif à la première grossesse et faible nombre d'enfants), l'utilisation d'hormones exogènes (contraceptifs oraux et hormonothérapie substitutive), l'alimentation (consommation d'alcool) et les caractéristiques anthropométriques (poids élevé, prise de poids à l'âge adulte et répartition de la graisse corporelle). D'autre part, l'allaitement maternel et l'exercice physique sont reconnus comme des facteurs de protection [16].

1.2.3 Classification des tumeurs

Le cancer du sein peut être classé selon différents stades en fonction de son étendue, soit du degré d'envahissement de l'organisme du patient par les cellules cancéreuses.

Cette classification est indispensable à la mise en œuvre d'un protocole de traitement adapté au cas par cas à chaque patiente et à chaque type de tumeur mammaire.

De fait, la personnalisation des protocoles de traitements anticancéreux est un aspect essentiel de leur succès. En fonction de son stade d'évolution au moment du diagnostic, le cancer du sein ne répond pas aux mêmes thérapies et ne présente pas le même pronostic. Les différents stades du cancer du sein illustré dans la Figure 1.2 correspondent à son degré d'évolution au moment de son diagnostic. Il existe différentes façons de classifier les tumeurs mammaires, la plus communément utilisée étant la classification TNM.

Celle-ci consiste à définir le stade de chaque tumeur cancéreuse en fonction de trois critères : la taille de la tumeur (T), les atteintes ganglionnaires (N) et la présence de métastases (M) [37].

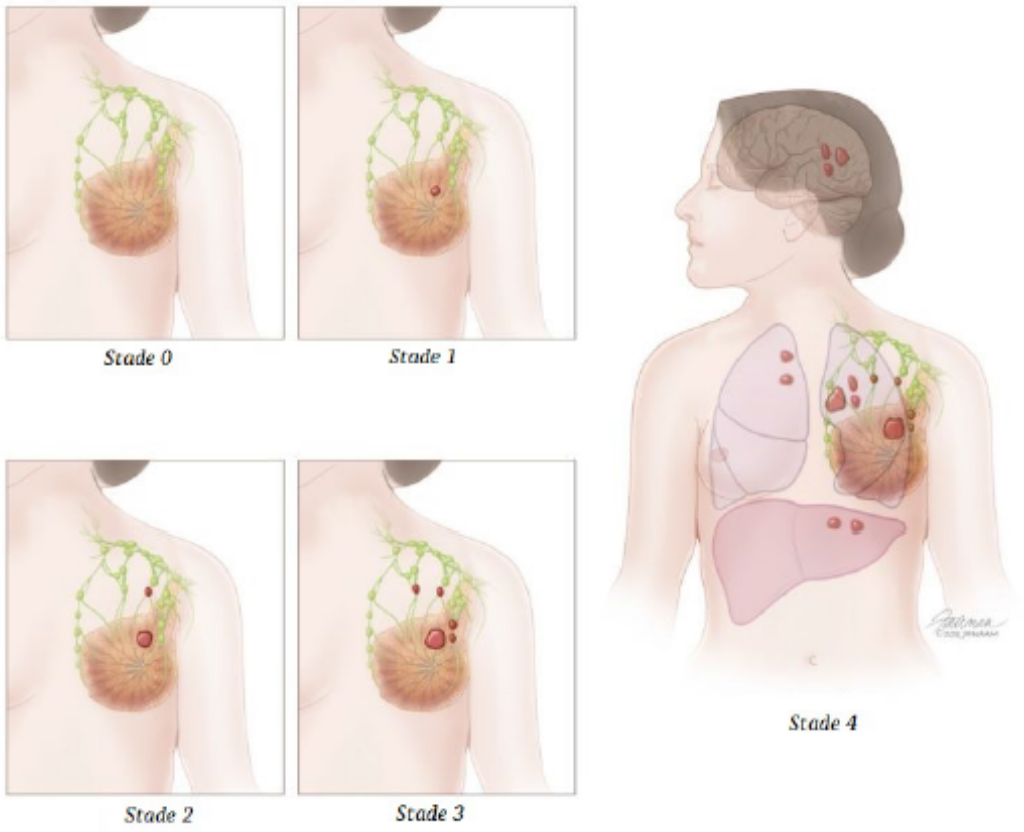
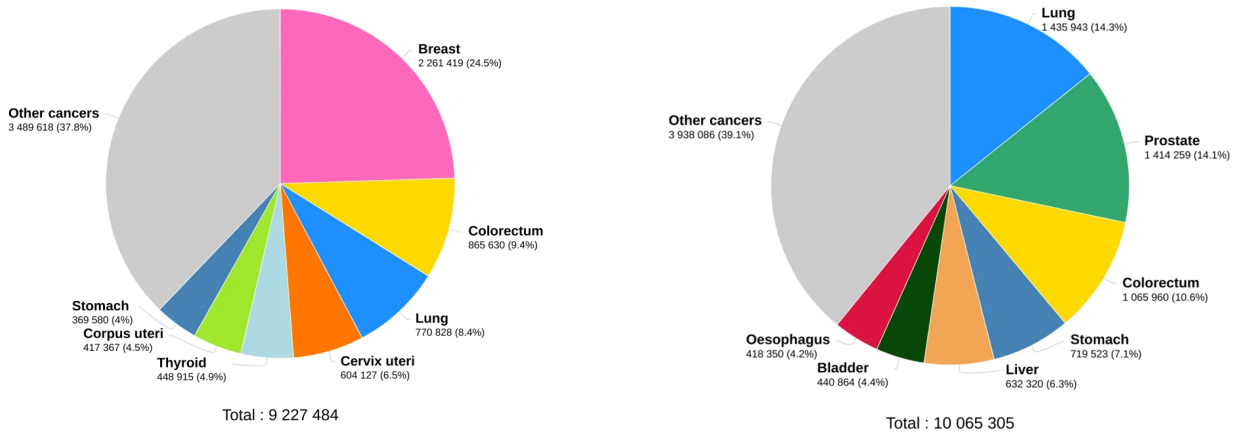


FIGURE 1.2 – Illustration des stades des tumeurs mammaires selon le système TNM [8]

1.2.4 Statistiques sur le cancer du sein

Les données chiffrées présentées ici proviennent du site de l'Organisation Mondiale de la Santé [3] sur l'estimation de l'incidence et de la mortalité par cancer dans le monde en 2020. L'OMS a compté près de 10 millions de décès dans la même année, soit presque un décès sur six, le cancer est l'une des principales causes de mortalité dans le monde, en 2020. Les cancers les plus courants (en termes du nombre de cas recensés) étaient les suivants : le cancer du sein (2,26 millions de cas), le cancer du poumon (2,21 millions de cas), le cancer colorectal (1,93 million de cas), le cancer de la prostate (1,41 million de cas), le cancer de la peau (1,20 million de cas) et le cancer de l'estomac (1,09 million de cas). Le cancer du sein est considéré comme le cancer le plus courant en termes d'incidence chez les femmes. En 2020 on a recensé 2,26 millions de cas féminins et 685 000 décès dus au cancer du sein dans le monde. À la fin de la même année, 7,8 millions de femmes en vie s'étaient vues diagnostiquer un cancer du sein au cours des cinq années précédentes, ce qui en fait le type de cancer le plus courant à l'échelle du globe. Le sexe féminin est le principal facteur de risque de cancer du sein. Cependant, environ 0,5 à 1 % des cas de cancer du sein surviennent chez les hommes. Dans de tels cas, le traitement suit les mêmes principes que chez les femmes. Présent dans tous les pays, le cancer du sein touche les femmes de tous âges à partir de la puberté, mais son incidence croît à mesure que l'âge avance. La mortalité par cancer du sein a peu évolué entre les années 1930 et les années 1970, période pendant laquelle la chirurgie était le mode primaire exclusif de traitement (mastectomie radicale). Le taux de survie a commencé à s'améliorer pendant les années 1990 lorsque des pays ont mis en œuvre des programmes de détection précoce associés à des programmes de traitement complets, avec des thérapies médicales efficaces [3]. Les Figures 1.3(a) et 1.3(b) présentent des diagrammes circulaires illustrant les taux d'incidence de différents type de cancer chez les femmes et les hommes, respectivement.



(a) Taux d'incidence du cancer chez les femmes.

(b) Taux d'incidence du cancer chez les hommes.

FIGURE 1.3 – Diagramme circulaire du taux d'incidence du cancer dans le monde en 2020 [4].

1.3 Imagerie biomédicale pour le dépistage du cancer du sein

Grâce aux progrès technologiques, l'imagerie médicale permet d'explorer l'intérieur du corps sans incision. Elle aide à prévenir les maladies et elle est considérée comme la méthode la plus efficace pour la détection précoce des tumeurs. Cependant, il est important de choisir la bonne technique d'imagerie en fonction des organes à examiner [18]. Ci-dessous une description des principales techniques d'imagerie employées en radiologie pour examiner les seins et détecter les anomalies à savoir la mammographie, l'échographie, la tomographie par ordinateur, la tomographie par émission de positrons et l'imagerie par résonance magnétique (IRM) sont présentées.

1.3.1 Mammographie

La mammographie est une modalité de dépistage qui utilise des rayons X à faible énergie pour examiner le sein, souvent en projetant les tissus en une image 2D [55]. Le dépistage par mammographie est recommandé pour les femmes par l'Organisation Mondiale de la Santé (OMS,2014), qui peut permettre un diagnostic précoce et améliorer le pronostic pour les patients potentiels [53]. Par ailleurs, la mammographie présente certaines faiblesses, tel que les rayonnement ionisants qui pourraient provoquer la formation des tumeurs à long terme. En outre, la qualité des images obtenues est souvent inférieure par rapport aux autres modalités d'imagerie. De plus elle ne parviens pas à détecter les anomalies dans les seins dense chez les femmes de moins de 35 ans [44].

1.3.2 Échographie

L'imagerie échographique est également une modalité de dépistage non invasive qui s'appuie sur l'emission des ondes ultrasonores avec des fréquences oscillant entre 2 et 20 MHz [16], pour visualiser l'intérieur du corps sans aucun rayonnement ionisants rendant son utilisations plus sûre quant aux populations sensibles comme les femmes enceintes. Son principal inconvénient réside dans sa capacité limitée à visualiser les structures profondes qui sont perturber par des obstacles tels que l'air ou l'os, bien que l'échographie soit une technique d'imagerie couramment utilisée pour ça polyvalence, sa qualité et son efficacité dépendent fortement des compétences de l'opérateur et des limitations inhérentes à la méthode [42]. Un gel de couplage entre la sonde et le sein est utilisé pour remédier à l'effet de refraction de l'air.

1.3.3 Tomodensitométrie

Appeler aussi tomographie par ordinateur, similaire a la mammographie, la tomodensitométrie (TDM) utilise des rayonnements ionisants. Toutefois, elle offre la capacité de créer la tomographie ou les plans du sein des patientes, en combinant les rayons X et des ordinateurs pour produire

des images détaillées des structures internes du corps, procurant ainsi une représentation en trois dimensions de l'organe [39]. En revanche la tomographie par ordinateur présente peu d'avantage dans le processus du diagnostique mammaire pour son incapacité à fournir assez d'informations pertinentes par rapport à la mammographie [16].

1.3.4 Tomographie par émission de positrons

La tomographie par émission de positrons (TEP) est une méthode d'imagerie médicale nucléaire qui implique l'injection d'une petite quantité de substance radioactive liquide dans le corps. Le plus souvent, une substance radioactive à base de sucre est administrée directement dans la circulation sanguine lors de l'examen TEP. Cette substance se concentre dans le corps, émettant des rayons gamma qui sont ensuite détectés par un scanner TEP [20]. Les données ainsi recueillies sont transformées en images détaillées par un ordinateur, ce qui permet d'observer le fonctionnement des tissus et des organes. En raison du métabolisme énergétique spécifique des cellules tumorales, elles consomment davantage de sucre pour leur croissance. La TEP permet donc de repérer ces tumeurs en surveillant l'utilisation du sucre dans le corps. Contrairement aux autres modalités d'imagerie axées sur les caractéristiques morphologiques, la TEP fournit des informations physiologiques sur la tumeur étudiée [63].

1.3.5 Imagerie par résonance magnétique

l'imagerie par résonance magnétique présentée dans la figure 1.4 est une modalité d'imagerie médicale totalement non invasive, son principes de fonctionnement repose sur l'exploitation des ondes radio et les champs magnétiques pour générer des informations détaillées, souvent sous forme d'une image 3D de l'intérieure de l'organe. Depuis son invention en 1971, après plusieurs test cliniques l'IRM a montrer sa polyvalence et son efficacité dans l'imagerie radiologique notamment pour la detection du cancer du sein [51]. De nos jours, les examens par IRM deviennent les modalités de balayage principales pour surveiller la réponse au traitement et la récurrence du cycle, offrant plus de détails sur les seins sans introduire de radiations ionisants [48].

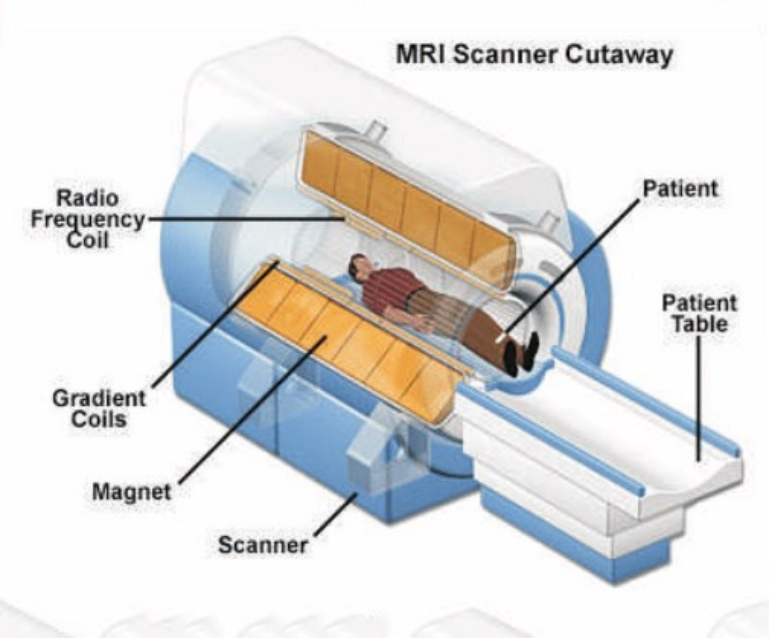


FIGURE 1.4 – Représentation d'un scanner IRM moderne [9].

1.4 Traitement du cancer du sein

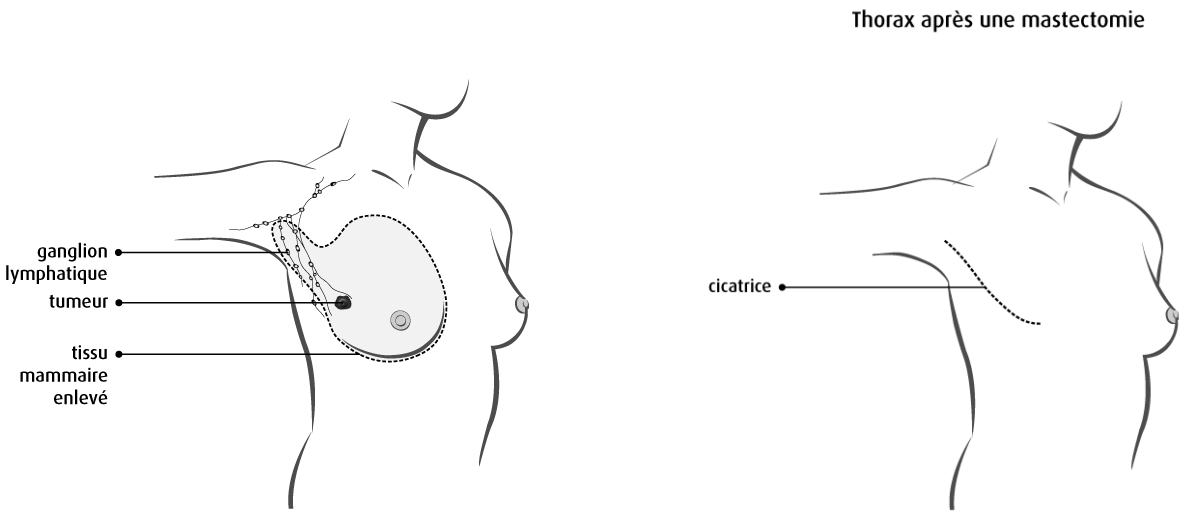
la prise en charge thérapeutique d'un cancer du sein doit être discutée et validée par une réunion pluridisciplinaire avant d'être expliqué à la patiente au cours de la consultation d'annonce la prise en charge d'un cancer du sein consiste d'une part à traiter localement le cancer d'autre part à traiter précocement par un traitement adjuvant les malades identifiés comme étant à risque métastatique [21].

1.4.1 Tumorectomie (mastectomie partielle)

Elle repose sur l'exérèse tumorale complète c'est à dire passant en tissu sein c'est une chirurgie conservatrice du sein dans laquelle le chirurgien extrait la tumeur du sein ainsi que du tissu moral autour de la tumeur. Un traitement néoadjuvant (chimio ou radiothérapie) est nécessaire après tumorectomie pour s'assurer de l'élimination de toutes les cellules cancéreuses microscopiques qui pouvaient toucher le tissu sous-jacent [22].

1.4.2 Mastectomie

C'est un traitement radical qui consiste en l'excision ou l'exérèse chirurgicale de la totalité de la glande mammaire en conservant le muscle grand pectoral tel qu'elle est présentée dans les Figures 1.5(a) et 1.5(b). Une chimiothérapie préopératoire, dite néodjuvante peut être proposé pour permettre une réduction tumorale avant son exérèse total [23].



(a) Avant (b) Après

FIGURE 1.5 – Illustration de la mastectomie radicale[5].

1.4.3 Chimiothérapie

La chimiothérapie désigne les traitements médicamenteux ayant pour but la destruction des cellules cancéreuses par des mécanismes non spécifiques. Trois cas principaux où la chimiothérapie peut être utilisée [24] :

1. Avant la chirurgie (thérapie néo-adjuvante) : Dans ce cas, la chimiothérapie est utilisée soit pour ralentir le développement d'une tumeur, soit pour réduire sa taille avant la tumorectomie.
2. Après la chirurgie (traitement adjuvant) : Utilisé pour réduire le risque de récidive du cancer.
3. En cas de métastases : Pour détruire les cellules cancéreuses qui peuvent s'être propagées à d'autres organes du corps.

1.5 Analyse d'images : principe et méthodes d'apprentissage automatique

Dans cette section, nous aborderons les principes et les méthodes d'apprentissage automatique appliqués à l'analyse d'images. Nous examinerons comment ces techniques sont utilisées pour extraire des informations significatives à partir d'images numériques.

1.5.1 Analyse d'images

L'analyse d'image à pour objectif d'extraire des informations significative à partir d'une image en fonction des objectifs et des besoins spécifiques de l'application. L'analyse d'image peut être

divisée en deux niveaux : analyse d'image de bas niveau et analyse d'image de haut niveau. Les algorithmes de bas niveau étudient les relations entre les valeurs numériques des pixels de l'image sans se soucier de la signification que les valeurs analysées ont dans la réalité [45]. A l'inverse, les algorithmes de haut niveau traitent les informations obtenues après l'application d'algorithmes de bas niveau pour décrire ou interpréter symboliquement le contenu de l'image [56].

1.5.2 Méthodes d'apprentissage automatique

L'*Apprentissage Automatique* (en anglais *Machine Learning (ML)*) est une méthode utilisée en intelligence artificielle qui consiste en un apprentissage statistique où chaque instance dans une base de données est décrite par un ensemble de caractéristiques. L'objectif principal est de reproduire une relation grâce à une fonction f, identifiée par un algorithme d'apprentissage, pour arriver à une décision ou prédiction Y en utilisant un jeu de données X. Les algorithmes de Machine Learning sont classés en trois catégories principales : *supervisés, non-supervisés, semi-supervisé* et *par renforcement* [26].

Apprentissage supervisé : Dans l'apprentissage supervisé (en anglais *Supervised Learning (SL)*), chaque exemple de données est associé à une étiquette ou une cible, et l'objectif est d'apprendre à prédire cette cible à partir des caractéristiques fournies. Les données d'apprentissage pour cette méthode comprennent à la fois l'échantillon d'entrée et la sortie désirée, qui peut être soit discrète/catégorique, soit réelle. Par exemple, dans le traitement d'images, un système d'intelligence artificielle peut être entraîné avec des images étiquetées pour classer les images non étiquetées dans différentes catégories (chacune dans sa catégorie). Parmi les algorithmes populaires d'apprentissage supervisé, on trouve la *Régression Linéaire*, la *Régression Logistique*, l'*Arbre de Décision*, les Machines à Vecteurs de Support, les Réseaux de Neurones, ainsi que les modèles non paramétriques comme les K-plus Proches Voisins, il est utilisé dans la classification d'e-mails en spam ou non-spam, la prédiction des prix immobiliers basée sur les caractéristiques des maisons, et la prévision de la demande future basée sur des données historiques de vente, il est utilisé dans la médecine, le transport logistique et les finance.

Apprentissage non-supervisé : L'apprentissage non-supervisé (en anglais *UNSupervised Learning (UNSL)*), également appelé apprentissage sans supervision, est une technique d'apprentissage automatique qui consiste à former des modèles sans étiqueter manuellement ou automatiquement les données précédemment. Grâce aux algorithmes utilisés, les données sont regroupées en fonction de leurs similitudes sans aucune interaction humaine permettant ainsi à un système d'intelligence artificielle de réagir à des informations qui ne sont pas classifiées ou étiquetées. C'est une méthode exploratoire pour la découverte des structures ou caractéristiques communes entre ces données leur offrant par ailleurs un moyen de voir les informations et aussi le cas échéant tester l'intelligence artificielle. Les systèmes de l'UNSL

sont souvent associés aux modèles d'apprentissages génératifs et utilisés dans plusieurs domaines tels que les chatbots, les véhicules autonomes, la reconnaissance faciale, les systèmes experts et les robots [6, 7].

Apprentissage semi-supervisé : L'*Apprentissage Semi-Supervisé* (en anglais *Semi-supervised learning* (SSL)) est une technique d'apprentissage automatique. Il se situe à mi-chemin entre l'apprentissage supervisé et non-supervisé, c'est-à-dire que l'ensemble des données est partiellement étiqueté. L'objectif principal du ASS est de surmonter les inconvénients à la fois de l'apprentissage supervisé et de l'apprentissage non-supervisé. L'apprentissage supervisé nécessite une grande quantité de données d'entraînement pour classer les données de test, ce qui est un processus coûteux et chronophage. D'autre part, l'apprentissage non-supervisé ne nécessite aucune donnée étiquetée, ce qui regroupe les données en fonction de la similarité entre les points de données en utilisant soit un regroupement, soit une approche de vraisemblance maximale. Le principal inconvénient de cette approche est qu'elle ne peut pas regrouper avec précision des données inconnues. Pour surmonter ces problèmes, la communauté de recherche a proposé SSL, qui peut apprendre avec une petite quantité de données d'entraînement et étiqueter les données inconnues (ou de test). SSL construit un modèle avec quelques motifs étiquetés comme données d'entraînement et traite le reste des motifs comme des données de test [57], il est utilisé dans divers domaines tel que le traitement du langages naturelles, la vision par ordinateur, la detection d'anomalies.

Apprentissage par renforcement : L'Apprentissage par Renforcement (en anglais *Reinforcement Learning* (RL)) fait référence à une classe de problèmes en apprentissage automatique dans lesquels un agent explore en toute autonomie un environnement et interagit avec ce dernier de sorte à percevoir des informations sur son état actuel et exécute des actions [31]. En retour, l'environnement fournit un signal de récompense (négatif ou positif), l'apprentissage par renforcement suit un processus itératif avec plusieurs étapes clés : État initial de l'environnement, Choix d'action en fonction de la politique, Réaction de l'environnement, Récompense, Mise à jour de la politique, Itération continue. Contrairement à d'autres méthodes d'apprentissage, l'agent ne reçoit aucune instruction explicite sur la manière de résoudre la tâche assignée. L'agent a pour objectif principal de maximiser le signal de récompense cumulatif tout au long de son interaction. RL est utilisé dans différents domaines tels que les jeux videos, la santé et la gestion des ressources [10].

Remarque 1.5.1.

- L'apprentissage supervisé vise à prédire une valeur cible à partir des caractéristiques fournies tandis que l'apprentissage non-supervisé vise à apprendre la distribution de probabilités des données ou certaines de ses propriétés, les techniques d'apprentissage automatique peuvent être utilisées pour les deux types de tâches.
- L'apprentissage supervisé est utilisé pour prédire des sorties à partir des entrées (données) étiquetées, l'apprentissage non-supervisé est utilisé pour découvrir et extraire des structures

cachées dans des données non étiquetées, tandis que l'apprentissage par renforcement est utilisé pour apprendre à interagir avec un environnement dynamique pour atteindre un objectif spécifique pour lequel il a été conçu.

- L'apprentissage semi-supervisé est particulièrement utile lorsque le marquage complet des données est coûteux ou difficile, offrant ainsi une solution efficace pour des problèmes où les données étiquetées sont limitées.
- L'apprentissage par renforcement implique un agent qui interagit avec un environnement dynamique et apprend à prendre des décisions séquentielles en s'ajustant à travers ses erreur pour accumuler une récompense maximale finale.

1.6 Conclusion

Le chapitre introductif fournit un aperçu essentiel du cancer du sein et de l'analyse des images, deux aspects essentiels de la lutte contre cette maladie dévastatrice. En mettant l'accent sur les progrès technologiques tels que l'imagerie médicale et l'intelligence artificielle, il met l'accent sur l'importance croissante de ces domaines pour améliorer le dépistage et la prise en charge du cancer du sein. En détail, l'anatomie et l'épidémiologie du cancer du sein sont abordées, offrant une compréhension approfondie de ses aspects fondamentaux. De plus, les statistiques sur l'incidence et la classification des tumeurs offrent des perspectives précieuses sur la portée et la gravité de cette maladie. Le chapitre met en lumière les avancées technologiques qui révolutionnent la chirurgie, la chimiothérapie et les méthodes d'apprentissage automatique.

Chapitre 2

État de l'art sur les approches de prédiction et classification du cancer du sein

Sommaire

2.1 Introduction aux méthodes d'apprentissage pour la détection et la classification du cancer du sein

Le cancer du sein demeure l'une des principales causes de décès chez les femmes à travers le monde. Le dépistage précoce et la classification précise des tumeurs mammaires jouent un rôle crucial dans la gestion et le traitement efficaces de cette maladie. Ces dernières années, les avancées dans le domaine de l'apprentissage automatique ont ouvert de nouvelles perspectives dans le domaine de la santé, offrant des outils puissants pour la détection et la classification du cancer du sein. Ce chapitre passe en revue des travaux présentés dans la littérature dans le cadre de la detection et de la classification du cancer du sein.

Dans cette optique, le chapitre est structuré de la manière suivante : la Section 2.3 présente une analyse des travaux fondés sur des méthodes d'apprentissage supervisé. La Section 2.4 examine les travaux basés sur des approches d'apprentissage non supervisé. Dans la Section 2.5, nous répertorions les recherches utilisant des méthodes d'apprentissage semi-supervisé. Par la suite, la Section 2.6 détaille les travaux exploitant les approches d'apprentissage par renforcement, puis nous découvrirons dans la Section 2.7 des travaux basés sur des approches d'apprentissage combinées. De plus, la Section 2.8 propose une comparaison des différentes études examinées. Avant de conclure, la Section 2.9 met en lumière quelques remarques et perspectives intéressantes à explorer à l'avenir. Enfin, nous concluons dans la Section 2.10.

2.2 Classification des approches de detection et de classification du cancer du sein

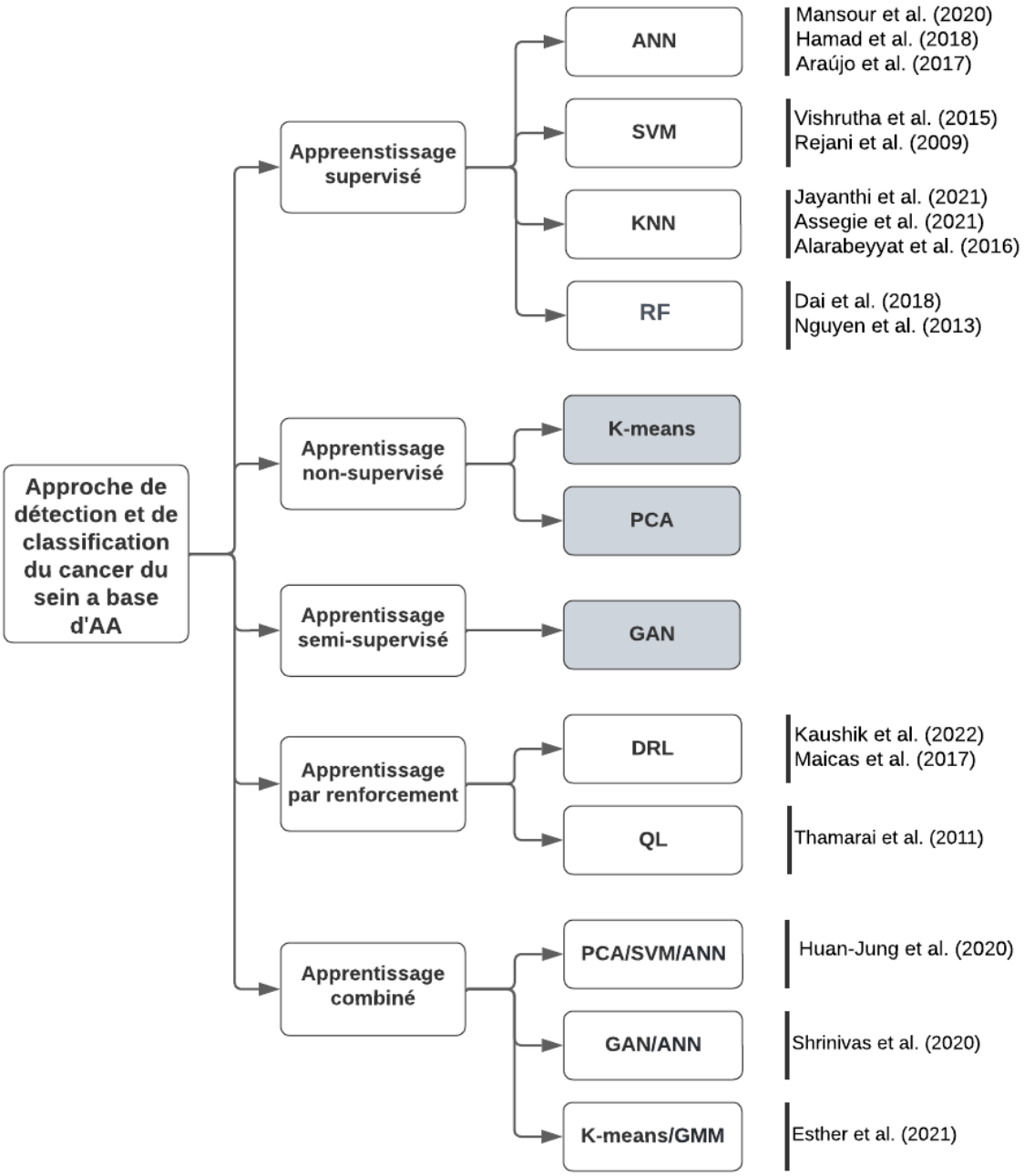


FIGURE 2.1 – Classification des Approches avec leurs travaux correspondants

2.3 Méthodes d'apprentissage supervisé

Dans cette section, nous présentons les travaux essentiels de la littérature dédiés à detection et classification du cancer du sein, se basant sur des méthodes d'apprentissage supervisé.

2.3.1 ANN

Les réseaux de neurones artificiels en anglais *Artificial Neural Networks* ANN sont des modèles informatiques inspirés du fonctionnement du cerveau humain. Ils sont utilisés pour résoudre une variété de tâches d'apprentissage automatique en apprenant à partir de données. Ces modèles sont composés de plusieurs couches de neurones interconnectés, permettant la capture de motifs complexes dans les données [34]. Mansour et al. [52] a utilisé un reseau de neurones convolutifs profonds pour la détection du cancer du sein, poursuivant cette approche Hamad et al. [36] proposent une méthode d'apprentissage profond qui incorpore des fonctions de suppression du bruit pour améliorer les caractéristiques des images médicales,atteignant une précision notable de 90%. Enfin, Araújo et al. [14] ont utilisé des réseaux de neurones convolutifs pour classer les images histologiques du cancer du sein en plusieurs catégories. Ces travaux soulignent l'efficacité des techniques d'apprentissage profond et des réseaux de neurones pour la détection et la classification du cancer du sein.

2.3.2 SVM

Les machines à vecteurs de support en anglais *Support Vector Machines* SVM sont des modèles d'apprentissage supervisé utilisés pour la classification et la régression. Ils fonctionnent en identifiant l'hyperplan qui sépare au mieux les différentes classes dans l'espace des caractéristiques[30]. Rejani et al. [58] mettent en avant l'utilisation d'un algorithme pour la détection de tumeurs à partir de mammographies. Ils soulignent l'importance de l'amélioration de la qualité de l'image, de la segmentation des images de mammographie et de l'extraction de caractéristiques pour améliorer la détection et le diagnostic du cancer du sein. Les tests ont montré que la méthode est efficace. De plus, l'étude de Vishrutha et al. [61] utilise l'algorithme SVM pour distinguer les tumeurs malignes des bénignes. Leur travail démontre l'efficacité des techniques de Machine Learning pour les diagnostics précoces du cancer du sein, surpassant même la précision des médecins expérimentés.

2.3.3 KNN

K plus Proches Voisins en anglais *K-Nearest Neighbors* KNN est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression. Il fonctionne en trouvant les k exemples d'entraînement les plus proches dans l'espace des caractéristiques et en utilisant leur étiquette pour prédire la classe ou la valeur de l'exemple de test [25]. Assegie et al. [15] proposent une méthode optimisée de détection du cancer du sein en utilisant l'algorithme KNN. L'ajustement des paramètres

améliore la précision, mais est coûteux en temps et en ressources. L'article de Alarabeyyat et al. [13] présente également une méthode de détection du cancer du sein en utilisant l'algorithme KNN. Enfin, l'article de Jayanthi et al. [43] utilise l'algorithme KNN pour la classification de la détection du cancer du sein avec le jeu de données du Wisconsin

2.3.4 RF

Forêt Aléatoire en anglais *Random Forest* RF est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression. Il fonctionne en construisant un ensemble de nombreux arbres de décision pendant l'entraînement et en fusionnant leurs prédictions pour obtenir une prédiction finale plus robuste [19]. Nguyen et al. [54] combinent la forêt aléatoire avec une sélection de caractéristiques, ce qui peut améliorer la précision du modèle en éliminant les caractéristiques non pertinentes, bien que le choix de la méthode de sélection des caractéristiques puisse influencer significativement les résultats. D'autre part, Dai et al. [27]. ont utilisé l'algorithme de la forêt aléatoire pour le diagnostic du cancer du sein, une approche robuste mais potentiellement difficile à interpréter et gourmande en temps de calcul.

2.4 Méthodes d'apprentissage non-supervisé

Ensuite, nous discutons de la façon dont les méthodes d'apprentissage non-supervisé sont utilisées pour la segmentation et la réduction de la dimensionnalité des données, ainsi que de leurs avantages et limitations dans le contexte de la détection du cancer du sein.

2.4.1 K-means

K-moyennes (en anglais *K-means*) est un algorithme de regroupement non-supervisé largement utilisé pour partitionner un ensemble de données en k groupes (clusters). Il fonctionne en itérativement assignant chaque point de données au cluster dont le centre (centroid) est le plus proche, puis en recalculant les centres de cluster en fonction des points attribués [49]. Cependant, aucune étude répertoriée dans la littérature n'a employé exclusivement le K-means pour la détection et la classification du cancer du sein.

2.4.2 PCA

L'Analyse en Composantes Principales en anglais *Principal Component Analysis* PCA est une méthode statistique qui permet de réduire la dimensionnalité des données en projetant celles-ci dans un nouvel espace orthogonal appelé espace des composantes principales. Cette projection permet de conserver le maximum de variance des données dans les premières composantes principales, ce qui facilite l'analyse et la visualisation des données [12]. Dans aucun des travaux référencés dans

la littérature, le PCA n'est utilisé en isolation, sans être combiné à d'autres méthodes, pour la détection et la classification du cancer du sein.

2.5 Méthodes d'apprentissage semi-supervisé

Dans cette section, nous examinons les principales études de recherche consacrées à la détection et à la classification du cancer du sein en utilisant des méthodes d'apprentissage demi-supervisé.

2.5.1 GAN

Les réseaux génératifs adversariaux en anglais *Generative Adversarial Networks* GAN sont une architecture de réseau de neurones profonds utilisée pour générer de nouvelles données réalistes. Ils se composent de deux réseaux neuronaux antagonistes : un générateur et un discriminateur. Le générateur tente de générer des données réalistes, tandis que le discriminateur tente de distinguer les données réelles des données générées. Cette compétition entre les deux réseaux entraîne une amélioration constante de la qualité des données générées [35]. Dans la littérature scientifique, aucune étude n'a exclusivement utilisé les GAN pour la détection et la classification du cancer du sein sans les combiner à d'autres modèles [28].

2.6 Méthodes d'apprentissage par renforcement

Dans cette section, nous passons en revue les travaux majeurs de la littérature portant sur la détection et la classification du cancer du sein à l'aide de méthodes d'apprentissage par renforcement.

2.6.1 DRL

L'apprentissage par renforcement profond en anglais *Deep Reinforcement Learning* DRL est une méthode d'apprentissage automatique qui vise à entraîner des agents à prendre des décisions en interagissant avec un environnement complexe. Ces agents utilisent des réseaux de neurones profonds pour apprendre une politique de décision basée sur des récompenses et des retours d'expérience obtenus par exploration active de l'environnement [47]. Maicas et al. [50]. ont utilisé l'apprentissage par renforcement profond DRL pour détecter activement les lésions mammaires à partir d'images DCE-MRI. Leur modèle s'adapte dynamiquement aux caractéristiques spécifiques des images médicales, mais l'entraînement de modèles DRL peut être complexe et nécessiter des ressources computationnelles importantes. En somme, cette approche offre un potentiel prometteur pour améliorer la détection précoce du cancer du sein. Kaushik et al. [46] ont utilisé l'apprentissage par renforcement profond DRL pour détecter activement les lésions mammaires à partir d'images

DCE-MRI. L'entraînement de modèles DRL peut être complexe et nécessiter des ressources computationnelles importantes. En somme, cette approche offre un potentiel prometteur pour améliorer la détection précoce du cancer du sein.

2.6.2 QL

Q-Learning QL est un algorithme d'apprentissage par renforcement utilisé pour apprendre une politique optimale dans un environnement basé sur les récompenses. Il fonctionne en itérativement mettre à jour les valeurs Q, qui représentent la récompense attendue pour prendre une action donnée dans un état donné, en utilisant une méthode de mise à jour basée sur la récompense reçue et l'estimation de la meilleure valeur Q future [62]. Thamarai et al. [60] présentent une méthode innovante de classification des tumeurs mammaires à partir de mammographies. Leur approche utilise l'algorithme d'apprentissage par renforcement Q-learning pour apprendre à classifier les tumeurs à partir des caractéristiques des images. Les points forts incluent l'utilisation novatrice de l'apprentissage par renforcement dans ce contexte, mais des limites peuvent résider dans le besoin de données volumineuses pour l'entraînement efficace et dans la complexité de mise en œuvre dans un environnement clinique.

2.7 Méthodes d'apprentissage combinées

Enfin, dans cette section nous examinons les principales études de la littérature concernant l'identification et la catégorisation du cancer du sein en utilisant des techniques d'apprentissage combinées.

2.7.1 PCA/SVM/ANN

Huan et al. [41] ont présenté une approche novatrice qui vise à détecter le cancer du sein en se basant sur neuf attributs individuels, tels que l'âge, l'indice de masse corporelle, le taux de glucose, l'insuline et une évaluation du modèle d'homéostasie. Cette méthode combine l'analyse en composantes principales (PCA) pour réduire la dimensionnalité des données, un réseau de perceptron multicouche pour extraire des caractéristiques pertinentes et une machine à vecteurs de support (SVM) par apprentissage par transfert pour la classification.

2.7.2 GAN/ANN

Desai et al. [59] dans leur étude sur la detection précoce du cancer du sein utilisent des méthodes d'apprentissage profond, malgré la rareté des images médicales étiquetées disponibles, ils appliquent les réseaux génératifs adversariaux (GAN) pour générer de nouvelles images de mammographie, augmentant ainsi les données pour compenser le manque d'images étiquetées. Ce qui améliorent les performances de detection de manière confédérale.

2.7.3 K-means/GMM

Esther et al. [32] ont utilisé deux approches de segmentation, le K-means et le modèle de mélange gaussien en anglais *Gaussian Mixture Model* GMM, pour segmenter différentes catégories d'images mammaires, telles que normales, bénignes et malignes. Ils ont démontré que leur approche hybride avait de meilleures mesures de performance, avec une précision de 95,5%.

2.8 Etude comparative et discussion

Dans le Tableau 2.2 suivant, nous résumons quelques travaux essentiels de la littérature dédiés à la detection et classification du cancer du sein en suivant la classification proposée, qui est donnée à la Figure 2.1, et en considérant des critères de comparaison bien choisis.

Classification		Critères / Articles	Modalité d'imagerie médicale	Objectif	Ensemble de données(Taille)	Inconvénients	Ac curacy
Type d'apprentissage	Approche adoptée						
Méthodes d'apprentissage supervisé	ANN	Hamad et al. (2018)	Mammographie	Classification du stade de tumeur du sein (bénigne, maligne ou normal)	DDSM mammographie (Kaggle)(2620)	Dépendance aux données spécifiques	90%
	SVM	Rejani et al. (2009)	Mammographie	Détection précoce du cancer du sein	mini-MIAS (322)	Sensibilité aux paramètres de prétraitement	88.75%
	KNN	Assegie et al. (2021)	Données tabulaires	Détection du cancer du sein	Wisconsin (Kaggle)(569)	Sensibilité au sur-ajustement des hyper paramètres	94.35%
	RF	Nguyen et al. (2018)	Données tabulaires	Classification des tumeurs du sein	Wisconsin University of California at Irvine (UCI)(569)	Risque de surajustement limitant la généralisation du modèle à d'autres contextes cliniques.	98,8%
Méthodes d'apprentissage non-supervisé	K-means						
	PCA						
Méthodes d'apprentissage semi-supervisé	GAN						
Méthodes d'apprentissage par renforcement	DRL	Maicas et al. (2017)	IRM dynamique	Détection et classification du cancer du sein	Utilisation du dataset public CBIS-DDSM(10239)	Discussion sur la sensibilité de l'algorithme aux variations de luminosité et de qualité d'image.	92%
	QL	Thamarai et al. (2011)	Données tabulaires	Classification du cancer du sein en deux classes (Bénin ou malin)	WBCD(569), WDBC(569) et WPBC(198) de l'UCI repository	Besoin de grandes quantités de données annotées.	98.90%
Méthodes d'apprentissage combinées	PCA/SVM/ ANN	Huan-Jung et al. (2020)	Données tabulaires	Détection et prédiction du cancer du sein	Manuel Gomes from the University Hospital Centre of Coimbra(116), Wisconsin(569)	Taille de l'ensemble de données limité et présente un déséquilibre entre les classes	86.97%
	GAN/ANN	Shrinivas et al. (2020)	Mammographie	Détection du cancer du sein	Dataset is downloaded from the DDSM(287)	Souffre du surapprentissage, et nécessite une validation soigneuse	87%
	K-means/GMM	Esther et al. (2021)	Mammographie	Détection et classification du cancer du sein(bénin, malin ou normal)	MIAS(322)	Forte dépendance a la qualité des données	95.5%

FIGURE 2.2 – Tableau comparatif des travaux de la littérature dédiés à la detection et classification du cancer du sein

- Les méthodes supervisées, telles que les SVM, ANN, KNN et RF, sont couramment utilisées

pour la détection et la classification du cancer du sein. Bien que prometteuses, elles ont des limites spécifiques à considérer.

Les SVM et les ANN sont connues pour leur dépendance à une grande quantité de données. L'approche adaptée par Hamad et al [36], utilisant des données de mammographie DDSM (2620), a obtenu un taux de précision de 90%, mais souffre de cette dépendance aux données spécifiques.

Les KNN, bien qu'efficaces, sont sensibles aux paramètres de prétraitement. L'approche de Assegie et al [15], utilisant l'ensemble de données Wisconsin (569), a obtenu un taux de précision de 94,35%, mais est sujette au surajustement des hyperparamètres.

Les RF, malgré leur performance élevée, peuvent manquer d'interprétabilité et présenter un risque de surajustement. L'approche de Nguyen et al [54], utilisant l'ensemble de données de l'University of California at Irvine (UCI) Wisconsin (569), a atteint un taux de précision de 98,8%, mais présente le risque de surajustement limitant la généralisation du modèle à d'autres contextes cliniques.

Il est également important de noter que le travail de Rejani et al [58] a donné lieu à un taux de précision relativement plus faible de 88,75%. Ceci est attribuable au fait que les jeux de données utilisés par les auteurs ne sont pas assez grands et à la sensibilité aux paramètres de prétraitement des SVM.

- Les méthodes d'apprentissage non-supervisé comme K-means et PCA offrent des alternatives pour segmenter et réduire la dimensionnalité des données liées au cancer du sein. K-means, largement utilisé pour partitionner les données en clusters, améliore la détection précoce du cancer du sein via une segmentation précise des images de mammographie. PCA, en projetant les données dans un nouvel espace appelé espace des composantes principales, facilite l'analyse des données. Ces méthodes ont le potentiel d'améliorer la détection et la classification du cancer du sein.
- Les méthodes semi-supervisées, comme les Réseaux Génératifs Antagonistes (GAN), offrent des alternatives intéressantes. Elles sont moins dépendantes des données étiquetées, mais peuvent être moins interprétables et générales.
- Les méthodes d'apprentissage par renforcement, telles que le DRL et le QL, offrent des approches novatrices pour la détection et la classification du cancer du sein. Dans cette approche, l'équipe de recherche dirigée par Maicas et al. [50] a utilisé le dataset public CBIS-DDSM de taille (10239) et a atteint un taux de précision de 92%. Cependant, un inconvénient significatif de cette méthode réside dans sa sensibilité aux variations de luminosité et de qualité d'image, ce qui pourrait limiter sa robustesse dans des environnements où les conditions d'éclairage et la qualité des images varient considérablement.

  Quant au QL, Thamarai et al. [60] ont utilisé les ensembles de données WBCD de taille (569), WDBC de taille (569) et WPBC de taille (198) de l'UCI repository, et ont obtenu

un taux de précision de 98,90%. Cependant, un inconvénient majeur de cette méthode est le besoin de grandes quantités de données annotées. Cette exigence en matière de données peut rendre son application difficile dans des domaines où les données annotées sont rares ou coûteuses à obtenir.

- Les méthodes d'apprentissage combinées, telles que PCA/SVM/ANN et GAN/ANN, sont largement utilisées dans la détection et la classification du cancer du sein. Cependant, chacune présente des avantages et des inconvénients spécifiques qui méritent d'être examinés.
  Pour l'approche adaptée PCA/SVM/ANN, présentée par Huan-Jung et al. [41], deux ensembles de données ont été utilisés : Manuel Gomes from the University Hospital Centre of Coimbra (116) et Wisconsin (699). Malgré un taux de précision honorable de 86,97%, cette méthode est limitée par la taille restreinte des ensembles de données, ce qui peut entraîner un déséquilibre entre les classes.
  En revanche, l'approche adaptée GAN/ANN, décrite par Shrinivas et al. [59], utilise un seul ensemble de données provenant du DDSM (287). Bien qu'elle affiche un taux de précision légèrement supérieur de 87%, elle présente des défis liés au surapprentissage, nécessitant ainsi une validation minutieuse pour garantir des résultats fiables.
  En examinant ces deux approches, il est clair que le choix entre elles dépendra de divers facteurs, notamment la disponibilité des données, la sensibilité au surapprentissage et le désir d'équilibrer les classes dans les ensembles de données. Ces éléments doivent être pris en compte lors du développement et de l'évaluation de méthodes d'apprentissage combinées pour la détection et la classification du cancer du sein.

2.9 Discussion et éventuels travaux futurs

L'examen approfondi des méthodes de détection et de classification du cancer du sein révèle plusieurs observations pertinentes ainsi que des pistes intéressantes pour des travaux futurs.

- **Complémentarité des approches supervisées et non-supervisées** : Les méthodes d'apprentissage supervisé comme les SVM, ANN, KNN et RF ont montré des performances prometteuses. Cependant, leur efficacité peut être limitée par la disponibilité et la qualité des données étiquetées. L'intégration des approches non-supervisées, comme K-means et PCA, pourrait améliorer la robustesse et la généralisation des modèles en exploitant des informations supplémentaires contenues dans les données non étiquetées.
- **Potentialité des méthodes semi-supervisées et par renforcement** :Les méthodes semi-supervisées, notamment les GAN, offrent des solutions intéressantes pour augmenter les données et améliorer les performances des modèles, surtout lorsque les données étiquetées sont rares. De plus, l'apprentissage par renforcement, bien que moins exploré, a démontré un potentiel prometteur pour la détection active et la classification des tumeurs mammaires. Cependant, des recherches supplémentaires sont nécessaires pour surmonter les défis liés à la

complexité et aux ressources computationnelles.

- **Approches d'apprentissage combinées** : L'utilisation de méthodes d'apprentissage combinées, comme PCA/SVM/ANN et GAN/ANN, a montré des résultats encourageants. Ces approches permettent de tirer parti des points forts de différentes techniques, mais nécessitent une validation rigoureuse pour éviter le surapprentissage et garantir des résultats généralisables. Des recherches futures pourraient explorer d'autres combinaisons de méthodes pour optimiser les performances de détection et de classification.
- **Création de nouveaux ensembles de données** : Un défi majeur dans la recherche sur le cancer du sein est le manque d'ensembles de données publics de grande taille et récents pour l'évaluation des modèles. Il serait crucial de constituer de nouveaux ensembles de données publics, comprenant un nombre significatif de cas, pour permettre une évaluation plus robuste des modèles proposés. Ces ensembles de données devraient inclure une diversité de types de tumeurs et de conditions cliniques pour améliorer la généralisation des modèles.
- **Importance de l'interprétabilité des modèles** :Bien que les modèles de deep learning, tels que les réseaux de neurones artificiels ANN, aient montré des performances élevées dans la détection et la classification des tumeurs, leur interprétabilité demeure un défi majeur. Il est donc essentiel de développer des méthodes d'explication des décisions prises par ces modèles, car cela peut renforcer la confiance des cliniciens et des patients dans les prédictions fournies. Une meilleure compréhension des mécanismes sous-jacents à ces décisions contribue non seulement à une adoption plus large des outils d'IA en milieu clinique, mais aussi à une prise de décision plus éclairée, ce qui est crucial pour le traitement des patients.
- **Optimisation des performances des modèles** : Pour maximiser l'efficacité des modèles, il est important de continuer à optimiser les hyperparamètres et à explorer des architectures de modèles novatrices. L'utilisation de techniques comme le transfert d'apprentissage, où des modèles pré-entraînés sur de grandes bases de données sont ajustés pour des tâches spécifiques de classification du cancer du sein, pourrait offrir des gains de performance significatifs.
- **Application clinique et intégration** : Un aspect crucial des futurs travaux sera de garantir que les modèles développés peuvent être intégrés de manière efficace dans les environnements cliniques. Cela implique une collaboration étroite avec des professionnels de la santé pour s'assurer que les outils d'IA sont adaptés aux flux de travail cliniques et qu'ils répondent aux besoins pratiques des cliniciens.

2.10 Conclusion

Ce chapitre a dressé un panorama des principales méthodes de détection et de classification du cancer du sein en utilisant des techniques d'apprentissage automatique. Les différentes approches, qu'elles soient supervisées, non supervisées, semi-supervisées, par renforcement ou combinées, ont été examinées en détail, mettant en lumière leurs avantages et leurs limites spécifiques.

À travers cette étude, il est apparu clairement que chaque méthode présente ses propres forces et faiblesses, soulignant ainsi la nécessité de poursuivre la recherche pour développer des approches plus robustes et généralisables. Les progrès continus dans le domaine de l'apprentissage automatique offrent des opportunités prometteuses pour améliorer la détection précoce et la classification précise des tumeurs mammaires.

Pour l'avenir, il serait judicieux d'explorer davantage les approches moins explorées et de poursuivre les efforts visant à constituer des ensembles de données plus vastes et diversifiés pour l'entraînement et l'évaluation des modèles. En outre, une collaboration étroite entre les chercheurs en informatique, les professionnels de la santé et les experts en imagerie médicale serait bénéfique pour garantir l'applicabilité et l'efficacité des méthodes développées dans des contextes cliniques réels.

Chapitre 3

Proposition d'une approche de classification du cancer du sein

Sommaire

3.1 Introduction

Les Réseaux de Neurones Artificiels (ANN) sont devenus des outils incontournables dans le domaine de l'apprentissage automatique, particulièrement pour les tâches de classification sur des données tabulaires. Les ANN peuvent travailler efficacement avec des données tabulaires en exploitant leur capacité à modéliser les interactions complexes entre les caractéristiques des données d'entrée.

Parallèlement, le clustering, domaine clé de l'apprentissage automatique, est largement étudié et comprend diverses méthodes pour aborder ce défi complexe. Les approches basées sur les K-means sont parmi les plus populaires, utilisant des stratégies telles que l'initialisation intelligente des centres de clusters et la réduction de la variance intra-cluster. De plus, la normalisation et la sélection appropriée des caractéristiques jouent un rôle essentiel en standardisant l'importance relative de chaque caractéristique dans le processus de regroupement. Cette intégration de la mise à l'échelle des caractéristiques aux méthodes K-means représente une avancée significative dans l'amélioration des techniques de clustering des données.

Dans ce chapitre, nous proposons une approche de classification du cancer du sein qui combine la méthode d'apprentissage profond ANN avec une méthode d'apprentissage automatique (RF ou K-means). Plus spécifiquement, nous utilisons en première phase les ANN pour classifier les tumeurs du sein en bénignes ou malignes en utilisant des données tabulaires du dataset *Wisconsin* [1]. Une fois cette classification initiale effectuée, nous appliquons en deuxième phase l'algorithme RF/K-means pour classifier les tumeurs malignes en trois stades distincts : précoce, intermédiaire et avancé. Cette approche hybride permet d'exploiter les forces des ANN pour la classification initiale et la méthode RF/K-means pour une classification plus détaillée des tumeurs malignes, offrant ainsi une évaluation plus complète et précise des stades du cancer mammaire. Par ailleurs, pour l'évaluation de la deuxième phase, nous avons proposé une méthode de complément du dataset *Wisconsin* par une nouvelle caractéristique qui représente les étiquettes des trois stades considérés.

Le chapitre est alors divisé comme suit : la Section 3.1 est consacrée à un rappel des méthodes d'apprentissage impliquées dans l'approche proposée pour la classification du cancer du sein. La Section 3.2 décrit les différentes étapes de notre approche. La Section 3.4.2 est dédiée à la méthode de complément du dataset *Wisconsin*. Enfin, on conclut le chapitre avec la Section 3.5..

3.2 Rappel sur les méthodes d'apprentissage ANN, RF et K-means

L'approche de classification du cancer du sein que nous proposions combine la méthode d'apprentissage profond ANN avec une méthode d'apprentissage automatique RF/K-means. Avant de donner la description des différentes étapes constituant l'approche proposée, nous avons jugé utile de faire un rappel des principes des méthodes combinées.

3.2.1 Méthode d'apprentissage ANN

Les Réseaux de Neurones Artificiels en anglais *Artificial Neural Networks* ANNs sont des systèmes d'apprentissage automatique inspirés du fonctionnement du cerveau humain. Un ANN est composé de plusieurs couches de nœuds appelés neurones artificiels, organisés en trois types de couches : l'entrée, les couches cachées et la sortie.

- **Couche d'entrée** : Cette couche reçoit les données d'entrée du réseau. Chaque neurone de cette couche correspond à une caractéristique ou une variable d'entrée.
- **Couches cachées** : Ces couches se situent entre l'entrée et la sortie. Elles effectuent des transformations non linéaires des données d'entrée grâce à des fonctions d'activation. Les neurones dans ces couches apprennent des représentations intermédiaires des données et peuvent détecter des caractéristiques complexes.
- **Couche de sortie** : Cette couche fournit la sortie du réseau, qui peut être une classification, une régression ou une autre forme de prédiction, selon la tâche que le réseau est formé pour accomplir.

Les neurones dans chaque couche sont interconnectés et chaque connexion possède un poids qui est ajusté pendant l'apprentissage. L'apprentissage se fait par propagation avant, où les entrées sont transmises à travers le réseau et par rétropropagation, où l'erreur entre la sortie prédite et la sortie réelle est utilisée pour ajuster les poids des connexions.

Les ANN peuvent être utilisés pour une variété de tâches telles que la reconnaissance d'images, la traduction automatique et la prévision des séries temporelles. Leur flexibilité et leur capacité à modéliser des relations complexes les rendent très puissants dans de nombreux domaines de l'intelligence artificielle [38].

3.2.2 Méthode d'apprentissage RF

Les forets aléatoires en anglais *Random Forest* RF est un algorithme d'apprentissage automatique supervisé, introduit par Leo Breiman en 2001, qui combine plusieurs arbres de décision pour améliorer les performances de classification ou de régression. Il est reconnu pour sa fiabilité, sa capacité à éviter le surapprentissage et à gérer de grands ensembles de données avec de nombreuses caractéristiques [19].

- **Processus de construction du modèle** :La construction d'un modèle Random Forest débute par le Bootstrap Aggregating (Bagging) : plusieurs ensembles de données bootstrap sont générés à partir de l'ensemble d'entraînement en échantillonnant avec remplacement. Chaque ensemble maintient la taille de l'échantillon original, introduisant ainsi des variations avec des échantillons répétés ou absents.

  Pour chaque ensemble bootstrap, un arbre de décision est construit. À chaque noeud de l'arbre, une sélection aléatoire de caractéristiques est utilisée pour déterminer la meilleure division, ce qui réduit la corrélation entre les arbres et augmente la diversité de la forêt.

Chaque arbre est développé sans élagage, c'est-à-dire qu'il se développe jusqu'à ce que chaque feuille soit pure ou qu'il n'y ait plus de divisions possibles.

- **Agrégation des prédictions** :En classification, chaque arbre prédit une classe et la classe finale prédite par le Random Forest est déterminée par un vote majoritaire des prédictions individuelles des arbres. En régression, chaque arbre donne une valeur de prédiction et la prédiction finale est la moyenne des prédictions des arbres.

3.2.3 Méthode d'apprentissage K-means

Comme nous avons brièvement déjà défini la méthode *K-means* dans dans la Sous-section 2.4.1 du Chapitre 2, *K-means* est un algorithme de regroupement non supervisé largement utilisé pour partitionner un ensemble de données en K groupes (clusters). Il fonctionne itérativement en assignant chaque point de données au cluster dont le centre (centroid) est le plus proche, puis en recalculant les centres des clusters en fonction des points attribués [49]. La méthode K-means fonctionne en suivant les étapes suivantes :

- *Initialisation :* Les centroids (points de référence) des clusters sont choisis aléatoirement parmi les données.
- *Assignation :* Chaque point de données est assigné au cluster dont le centroid est le plus proche.
- *Mise à jour :* Les centroids sont mis à jour en calculant la moyenne des points de données assignés à chaque cluster.
- *Répétition :* Les étapes 2 et 3 sont répétées jusqu'à ce que les centroids ne changent plus.

La méthode K-means est efficace pour identifier des patterns dans des données non étiquetées, mais elle présente des limitations, comme la dépendance aux initialisations et la sensibilité aux paramètres choisis [49].

3.3 Proposition d'une approche de classification du cancer du sein

Dans cette section, nous présentons une approche qui combine la méthode d'apprentissage profond (deep learning) ANN avec une méthode d'apprentissage automatique RF ou K-means en vue d'une classification du cancer du sein. L'approche en question est composée de quatre phases dont chacune est constituée d'un ensemble d'étapes.

3.3.1 Organigramme de l'approche proposée

L'organigramme de l'approche proposée, qui résume les quatre phases avec leurs différentes étapes, est représenté dans la Figure 3.1 suivante :

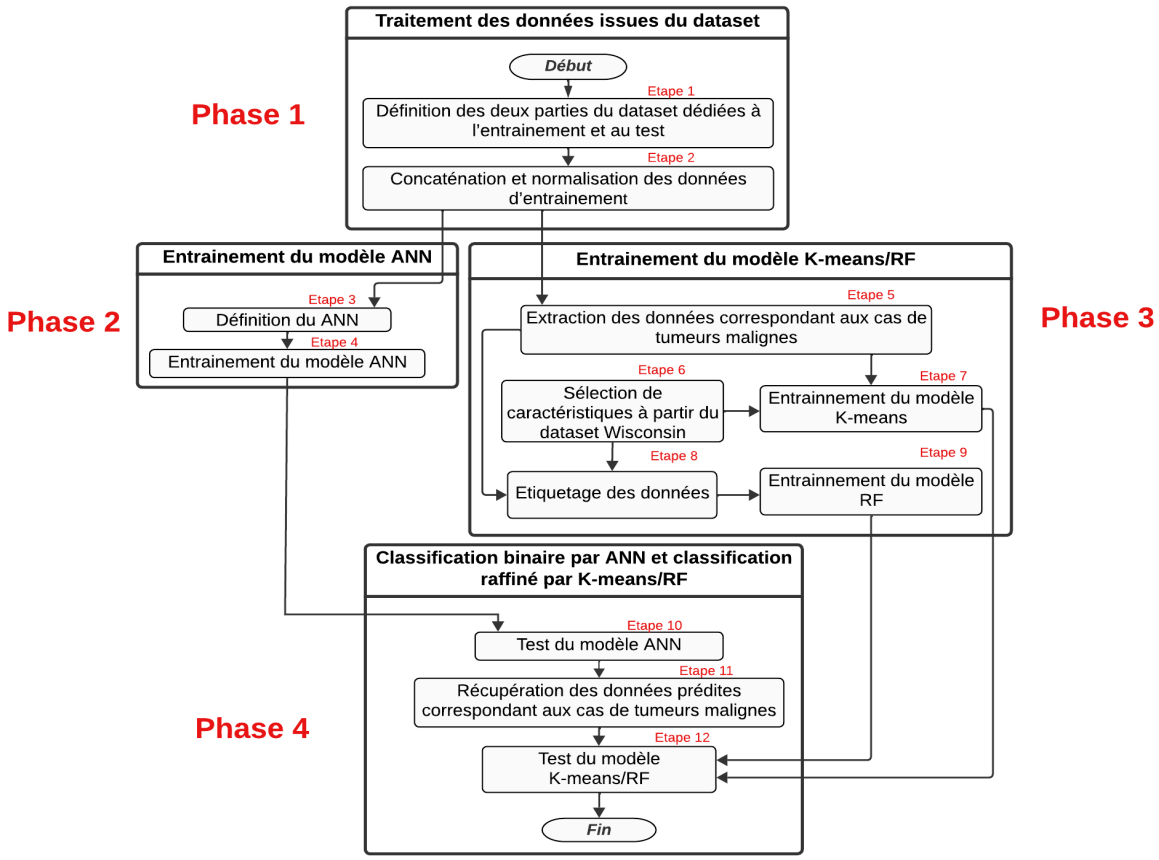


FIGURE 3.1 – Organigramme de l'approche proposée.

3.3.2 Description de l'approche proposée

Dans ce qui suit, nous présentons en détail les quatre phases ainsi que leurs différentes étapes qui constituent l'approche proposée.

Phase 01 : Traitement des données issus du dataset.

Étape 1 (Définition des deux parties du dataset dédiées à l'entraînement) : Les données sont lues à partir d'un fichier CSV qui contient les informations nécessaires pour l'analyse. Cela permet de manipuler les données de manière structurée en utilisant les bibliothèques pandas. La colonne 'diagnosis' contient des valeurs catégorielles "M" pour malin et "B" pour bénin). Ces valeurs sont mappées à des valeurs numériques (1 pour mali et 0 pour bénigne) afin de pouvoir les utiliser comme cible dans l'entraînement du modèle. Les données sont divisées en deux groupes distincts, en fonction de la valeur de la colonne 'diagnosis'. Cela permet de traiter séparément les exemples malins et bénins, facilitant ainsi la gestion des proportions dans l'ensemble d'entraînement. Pour chaque groupe (malin et bénin), 70% des données sont sélectionnées aléatoirement pour consti-

tuer l'ensemble d'entraînement (EE). Cela garantit que les proportions des classes sont maintenues dans l'ensemble d'entraînement. Pour la Concatenation et normalisation des données d'entraînement, aller à l'Étape 2.

Étape 2 (Concatenation et normalisation des données d'entraînement) : Les deux sous-ensembles d'entraînement, contenant respectivement les exemples malins et bénins, sont fusionnés pour former un ensemble d'entraînement complet. Cela permet de disposer d'un ensemble de données équilibré pour l'entraînement du modèle. L'ensemble d'entraînement est divisé en deux parties : la première partie (P_1) est celle où on considère toutes les colonnes sauf "diagnosis", "id", et 'Unnamed : "32", et la deuxième partie (P_2) est celle où on considère que la colonne cible "diagnosis". Cela prépare les données pour l'entraînement du modèle. Les caractéristiques de P_1 sont normalisées pour que leurs valeurs soient comprises entre 0 et 1. Cela est réalisé à l'aide du *MinMaxScaler* de *sklearn* du Language de programmation Python. La normalisation est importante car elle assure que toutes les caractéristiques soient de même poids lors de l'entraînement du modèle, ce qui améliore la convergence. Pour la définition de l'ANN, aller à l'Étape 3 et pour l'extraction des données correspondant aux cas de tumeurs malignes, allez à l'Étape 5.

Phase 02 : Entraînement du modèle ANN (en parallèle avec la phase 3).

Étape 3 (Définition du modèle ANN) : Un modèle de réseau de neurones séquentiel est défini. Il est composé de 04 couches principales : une couche d'entrée qui utilise les 30 caractéristiques de P_1 et P_2, une couche cachée comprenant 16 neurones avec une fonction d'activation ReLU permettant au modèle de capturer des relations non linéaires dans les données, une couche dropout avec un taux de 0.2 pour la régularisation, et une couche de sortie qui utilise une unité avec une activation sigmoïde pour produire une probabilité indiquant la classe prédite (malin ou bénin). Le modèle est compilé en utilisant l'optimiseur Adam, qui est efficace pour l'entraînement de réseaux de neurones. La précision est utilisée comme métrique pour évaluer la performance du modèle. Pour entraîner le modèle ANN, allez à l'Étape 4.

Étape 4 (Entraînement du ANN) : L'ensemble d'entraînement EE est partagé en deux parties : la première partie est constituée de 80 % des données (dédiées à l'entraînement) et la deuxième partie est constituée de 20 % des données (dédiées à la validation). Le modèle est entraîné sur les données normalisées. L'entraînement utilise un batch size de 64, ce qui signifie que les poids du modèle sont mis à jour après avoir vu 64 exemples des données d'entraînement. Le modèle est entraîné pendant 600 époques, ce qui donne au modèle suffisamment de temps pour apprendre les relations dans les données. Une fraction de 20% des données d'entraînement est utilisée pour la validation, permettant de surveiller la performance du modèle sur des données non vues pendant

l'entraînement. Pour effectuer le test du modèle ANN, allez à l'Étape 10.

Phase 03 : Entraînement du modèle K-means/RF (en parallèle avec la phase 2).

Étape 5 (Extraction des données correspondants aux cas de tumeurs malignes) : Dans cette étape, nous exploitons 70 % des données diagnostiquées comme "M" (malin), *i.e.* de la colonne "Diagnosis", identiques à la partie des cas malins utilisés pour l'entraînement du ANN. Toutefois, cette fois-ci, ces données sont employées pour l'entraînement de l'algorithme de partitionnement K-means. L'objectif principal de cette étape est de restreindre l'entraînement aux seuls cas malins en excluant les données du test, étant donné que notre intérêt porte spécifiquement sur la classification des stades uniquement pour les cas prédits comme malins. Pour l'entraînement du modèle K-means, allez à l'Étape 7.

Étape 6 (Selection de caractéristiques à partir du dataset) : Dans cette étape, nous appliquons une méthode de sélection de caractéristiques basée sur l'analyse de variance (ANOVA) pour identifier les caractéristiques les plus significatives dans un ensemble de données [11]. L'ANOVA compare chaque caractéristique à toutes les autres pour évaluer sa capacité à distinguer entre différents groupes, produisant une valeur F qui mesure la variance entre les groupes par rapport à la variance au sein des groupes. Les caractéristiques avec les valeurs F les plus élevées, indiquant une plus grande importance discriminante, sont ensuite sélectionnées. Ce processus permet de réduire la dimensionnalité des données en ne conservant que les caractéristiques les plus informatives. Pour l'entraînement du modèle K-means, allez à l'Étape 7, et pour étiqueter les données, allez à l'Étape 8.

Étape 7 (Entraînement du modèle K-means) : Dans ce fragment, une fonction est définie et prend en entrée les données d'entraînement, une liste de caractéristiques sélectionnées à utiliser pour l'entraînement, et le nombre de clusters à générer initialisé à trois (0 : précoce, 1 : intermédiaire, 2 : avancé). Les données sont d'abord mises à l'échelle pour garantir que toutes les caractéristiques ont la même échelle. Ensuite, l'algorithme K-means est appliqué aux données d'entraînement mises à l'échelle. Après avoir ajusté le modèle, les distances des centres de clusters sont calculées et les indices des clusters sont triés en fonction de ces distances. Une carte de correspondance est créée pour les clusters triés. Enfin, le modèle est utilisé pour prédire les clusters pour l'ensemble d'entraînement.Pour effectuer le test du modèle K-means/RF, allez à l'Étape 12.

Étape 8 (Étiquetage des données) : Pour entraîner le modèle Random Forest (RF), nous avons besoin d'étiqueter les données en suivant une méthode d'affectation des étiquettes qui sera présentée en détail à la Sous-section 3.4.2. Pour entraîner le modèle RF, aller à l'Étape 9.

Étape 9 (Entraînement du modèle RF) : Dans cette étape, l'entraînement du modèle Random Forest (RF) consiste à construire une multitude d'arbres de décision à partir de sous-échantillons des données d'entraînement précédemment étiquetées. Pour effectuer le test du modèle K-means/RF, allez à l'Étape 12.

Phase 04 : Classification binaire par ANN et classification raffiné par K-means/RF.

Étape 10 (Test du modèle ANN) : Après l'entraînement, le modèle doit être évalué sur un ensemble de test pour mesurer sa performance sur des données qu'il n'a jamais vues. Cette étape implique l'utilisation de métriques telles que l'accuracy, la précision, le F_1-Score pour évaluer la capacité du modèle à généraliser et à faire des prédictions correctes. Les données de test sont prétraitées de la même manière que les données d'entraînement (normalisation incluse) avant d'être passées au modèle pour prédiction. Pour la récupération des données prédites correspondant aux cas de tumeurs malignes, allez à l'Étape 11.

Étape 11 (Récupération des données prédites correspondant aux cas de tumeurs malignes) : Dans cette étape nous récupérons les données prédites durant la Phase 01 correspondants aux cas de tumeurs malignes. La raison pour laquelle nous avons effectué cette étape est de pouvoir tester le modèle K-means après son entraînement en se concentrant sur les cas les plus critiques, c'est-à-dire les tumeurs malignes, pour une analyse approfondie. En effet, notre objectif est non seulement de détecter la présence d'une tumeur maligne (cancer), mais aussi de déterminer le stade de la maladie. Pour ce faire, nous avons besoin de données plus détaillées sur les caractéristiques des tumeurs malignes. Puisque le modèle ANN utilisé dans la Phase 01 n'atteigne pas forcément une précision de 100%, il est important de noter que les données ainsi récupérées ne représentent pas tout à fait la totalité des cas malins figurant dans l'ensemble de test du ANN. Cette diminution est attribuable aux erreurs de prédiction du ANN sur les données de test, ce qui a entraîné une perte de certaines données malignes lors de la phase de prédiction. Pour effectuer le test du modèle K-means/RF, allez à l'Étape 12.

Étape 12 (Test du modèle K-means/RF) : Après avoir entraîné les modèles K-means et RF sur leurs ensembles d'entraînements correspondant, cete étape consiste à tester les modèles sur l'ensemble de test qui provient des résultats du test du ANN en Phase 01. Cette étape est essentielle pour évaluer la performance des deux modèles K-means/RF sur de nouvelles données non vues lors de l'entraînement.

3.4 Dataset *Wisconsin* : description et proposition pour son complément

Dans cette section, dans un premier temps, nous présentons le dataset *Wisconsin* tel qu'il existe dans la littérature et, dans un deuxième temps, pour des besoins d'évaluation de notre approche, nous proposons une méthode de complément du dataset par une nouvelle colonne nommée "*Stage*" (en Français "Stade") qui représente un raffinement de la colonne "Diagnosis" en trois stades : Précoce, Intermédiaire, et Avancé, pour les patientes présentant une tumeur maligne.

3.4.1 Description du dataset *Wisconsin*

Pour l'évaluation de l'approche proposée, nous avons besoin d'un dataset avec des valeurs tabulaires. Notre choix est porté sur le dataset public *Wisconsin* trouvé sur le site Kaggle [1]. Ce dernier est constitué de 569 lignes correspondant aux différentes patientes et 32 colonnes représentant les caractéristiques du dataset. Ces dernières sont définies comme suit :

1. **id** : Identifiant unique de chaque patiente ou observation.
2. **diagnosis** : Diagnostic de la tumeur, soit "M" pour maligne (cancéreuse) ou "B" pour bénigne (non cancéreuse).
3. **radius_mean** : Valeur moyenne du rayon de la tumeur.
4. **texture_mean** : Valeur moyenne de la texture de la tumeur.
5. **perimeter_mean** : Valeur moyenne du périmètre de la tumeur.
6. **area_mean** : Valeur moyenne de la surface de la tumeur.
7. **smoothness_mean** : Valeur moyenne de la régularité de la surface de la tumeur.
8. **compactness_mean** : Valeur moyenne de la compacité de la tumeur.
9. **concavity_mean** : Valeur moyenne de la concavité de la tumeur.
10. **concave_points_mean** : Valeur moyenne du nombre de points de concavité de la tumeur.
11. **symmetry_mean** : Valeur moyenne de la symétrie de la tumeur.
12. **fractal_dimension_mean** : Valeur moyenne de la dimension fractale de la tumeur.
13. **radius_se** : Erreur standard du rayon de la tumeur.
14. **texture_se** : Erreur standard de la texture de la tumeur.
15. **perimeter_se** : Erreur standard du périmètre de la tumeur.
16. **area_se** : Erreur standard de la surface de la tumeur.
17. **smoothness_se** : Erreur standard de la régularité de la surface de la tumeur.
18. **compactness_se** : Erreur standard de la compacité de la tumeur.

19. **concavity_se** : Erreur standard de la concavité de la tumeur.
20. **concave_points_se** : Erreur standard du nombre de points de concavité de la tumeur.
21. **symmetry_se** : Erreur standard de la symétrie de la tumeur.
22. **fractal_dimension_se** : Erreur standard de la dimension fractale de la tumeur.
23. **radius_worst** : Valeur maximale du rayon de la tumeur.
24. **texture_worst** : Valeur maximale de la texture de la tumeur.
25. **perimeter_worst** : Valeur maximale du périmètre de la tumeur.
26. **area_worst** : Valeur maximale de la surface de la tumeur.
27. **smoothness_worst** : Valeur maximale de la régularité de la surface de la tumeur.
28. **compactness_worst** : Valeur maximale de la compacité de la tumeur.
29. **concavity_worst** : Valeur maximale de la concavité de la tumeur.
30. **concave_points_worst** : Valeur maximale du nombre de points de concavité de la tumeur.
31. **symmetry_worst** : Valeur maximale de la symétrie de la tumeur.
32. **fractal_dimension_worst** : Valeur maximale de la dimension fractale de la tumeur.

Principalement, le dataset *Wisconsin* ne contient que les cas où la tumeur est Bénigne, représentée par l'étiquette "B", ou Maligne, représentée par l'étiquette "M", ce qui est donnée dans la colonne représentant la caractéristique "Diagnostic". C'est ce que nous avons besoin pour l'évaluation de la première phase de notre approche. Par ailleurs, pour effectuer une évaluation de la deuxième phase de l'approche proposée, nous avons besoin de compléter le dataset *Wisconsin* par une nouvelle colonne qui représente un raffinement des cas correspondant aux patientes qui présentent des tumeurs malignes. Ainsi, comme annoncé ci-dessus, nous avons utilisé trois stades de classification, à savoir : précoce, intermédiaire ou avancé.

Dans la sous-section suivante, nous présentons les détails de la méthode de complément du dataset utilisé.

3.4.2 Proposition d'une méthode de complément du dataset *Wisconsin*

Dans cette sous-section, nous proposons une méthode de complément du dataset *Wisconsin* par une colonne nommée "Stage", constituée de 03 étiquettes précoce, intermédiaire et avancé, en vue d'effectuer une évaluation de la deuxième phase de l'approche proposée. La méthode est constituée de 04 étapes présentées comme suit :

Étape 1 (Selection de caractéristiques) :

Pour raffiner la classification des tumeurs malignes, figurant dans la colonne "Diagnosis", en trois stades, précoce, intermédiaire, avancé, nous avons choisi 10 caractéristiques

$\{c_1, ..., c_{10}\}$ qui semblent les plus significatives du dataset, en utilisant un programme de sélection de caractéristiques "*SelectKBest*" avec le test ANOVA [11]. Plus spécifiquement, le programme sélectionne les caractéristiques en calculant les scores ANOVA pour chaque caractéristique, puis il choisit les "K" meilleures en fonction de ces scores (dans notre cas $K = 10$). Cela permet de garder les caractéristiques les plus pertinentes pour la classification. Ainsi, les caractéristiques sélectionnées pour le complément du dataset sont : `radius_mean`, `perimeter_mean`, `area_mean`, `concavity_mean`, `concave points_mean`, `radius_worst`, `perimeter_worst`, `area_worst`, `concavity_worst`, `concave points_worst`.

Étape 2 (Normalisation des données des caractéristiques sélectionnées) :

Nous appliquons une normalisation aux données des caractéristiques afin de les mettre sur une échelle commune. Cela se fait en utilisant la méthode *MinMaxScaler* pour transformer les données de sorte que les valeurs de chaque caractéristique soient comprises entre 0 et 1.

Étape 3 (Définition des intervalles associés aux stades précoce, intermédiaire et avancé en utilisant les 30e et 65e centiles) :

Étant donné que les caractéristiques ont été normalisées dans l'intervalle [0,1], les centiles (appelés aussi percentiles) seront les mêmes pour toutes les caractéristiques. En statistique descriptive, un centile (ou percentile) est une des 99 valeurs qui divisent une distribution de données en 100 parts égales de sorte que le p-ième centile soit la valeur supérieure à p % des autres valeurs. Les centiles sont un cas particulier des quantiles.

Dans notre cas, pour des besoins de classification d'une tumeur en l'un des trois stades précoce, intermédiaire ou avancé, nous avons considéré les 30e et 65e centiles. Ainsi, nous définissons trois intervalles de classification comme suit :

[0, 0.30] : l'intervalle correspondant aux valeurs de caractéristiques classées dans le stade "précoce".

]0.30, 0.65] : l'intervalle correspondant aux valeurs de caractéristiques classées dans le stade "intermédiaire".

]0.65, 1] : l'intervalle correspondant aux valeurs de caractéristiques classées dans le stade "avancé".

La Figure 3.2 résume les trois intervalles de classification définis ci-dessus :

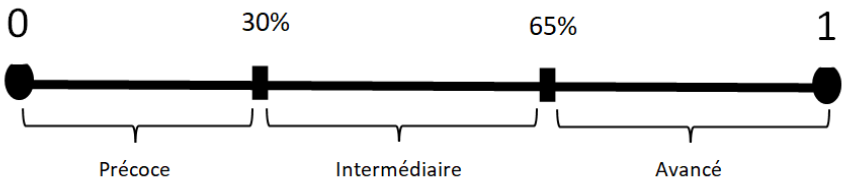


FIGURE 3.2 – Intervalles associés aux stades

Étape 4 (Classification et construction d'une nouvelle colonne "Stage") :

Pour classifier une patiente $\mathcal{P}$ présentant une tumeur maligne dans l'un des trois stades, nous procédons comme suit :

- Nous considérons les valeurs des 10 caractéristiques choisies associées à la patiente $\mathcal{P}$;
- Nous évaluons la position de chaque valeur des 10 caractéristiques par apport aux 30e et 65e centiles.
- Nous comptons le nombre de caractéristiques ayant leurs valeurs se situant dans chaque intervalle (précoce, intermédiaire, avancé).
- La patiente $\mathcal{P}$ est assignée au stade correspondant à l'intervalle contenant le plus de caractéristiques.
- Si le nombre maximal de caractéristiques figure dans deux intervalles, nous utilisons les distances aux 30e et/ou 65e centiles pour déterminer le stade :
  - ▷ Si la patiente $\mathcal{P}$ a le même nombre de caractéristiques dans les stades précoce et intermédiaire, nous calculons :
    - ✓ La somme des distances absolues aux 30e centile des valeurs des caractéristiques se trouvant dans le stade précoce.

$$D_{precoce} = \sum_{v \in [0,0.30]} |v - 0.30|,$$

où v est une valeur de caractéristique donnée de la patiente $\mathcal{P}$.

    - ✓ La somme des distances absolues aux 30e centile des valeurs des caractéristiques se trouvant dans le stade intermédiaire.

$$D_{intermediaire} = \sum_{v \in]0.30,0.65]} |v - 0.30|.$$

Ainsi, la patiente $\mathcal{P}$ est assignée au stade précoce si $D_{precose} > D_{intermediaire}$, sinon elle est assignée au stade intermédiaire.

  - ▷ Le même principe s'applique pour les égalités entre les stades intermédiaire et avancé en utilisant les distance suivantes par rapport au 65e centile :

$$D_{intermediaire} = \sum_{v \in]0.30,0.65]} |v - 0.65|,$$

$$D_{avance} = \sum_{v \in]0.65,1]} |v - 0.65|.$$

  - ▷ Le même principe s'applique pour les égalités entre les stades précoce et avancé en utilisant les distance suivantes par rapport au 30e et 65e centiles, respectivement :

$$D_{precoce} = \sum_{v \in [0,0.30]} |v - 0.30|,$$

$$D_{avance} = \sum_{v \in]0.65,1]} |v - 0.65|.$$

Après avoir classé chaque patiente dans sa catégorie ("Stage"), on associe la valeur "0" à chaque patiente classée dans le stade "précoce", la valeur "1" à chaque patiente classée dans le stade "intermédiaire" et la valeur "2" à chaque patiente classée dans le stade "avancé". Ainsi, nous avons défini une nouvelle colonne appelée "Stage", qui résume ces informations, représentant un raffinement de la colonne "diagnosis" pour détailler les tumeurs malignes selon les trois stades considéré.

3.5 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche pour classifier le cancer du sein en combinant une méthode d'apprentissage profond ANN avec une méthode d'apprentissage automatique RF/K-means.

Notre méthode se déroule en deux phases principales : d'abord, la classification des tumeurs en bénignes ou malignes avec un ANN, puis la classification des tumeurs malignes en trois stades (précoce, intermédiaire, avancé) avec *RF/K-means*. Cette combinaison exploite les avantages des deux techniques, offrant une solution plus précise pour évaluer les stades du cancer du sein. Les résultats montrent que l'intégration des ANN et de *RF/K-means* améliore la précision de la classification et fournit une évaluation plus détaillée de la progression du cancer, ce qui peut être utile pour le diagnostic et le traitement clinique. Le chapitre suivant sera consacré à l'évaluation des performances de l'approche proposée.

Chapitre 4

Chapitre 4. Évaluation de l'approche de classification du cancer du sein

Sommaire

4.1 Introduction

Dans ce chapitre, nous évaluons notre approche de classification des tumeurs du sein en deux types, bénin ou malin, ainsi que la classification par stades des tumeurs malignes.

Le chapitre est organisé comme suit : Dans la Section 4.2, nous présentons l'environnement de l'implémentation. Dans la Section 4.4, nous exposons les différentes métriques utilisées dans l'évaluation. Dans la Section 4.5, nous exhibons les résultats d'évaluation de l'approche proposée tout en effectuant des comparaisons appropriées par phase. Dans la Section 4.6, nous présentons les conclusions principales du chapitre.

4.2 Présentation de l'environnement de l'implémentation

L'implémentation de notre modèle a été réalisée en utilisant les outils et bibliothèques suivants :

- **Python** : Langage de programmation principal.
- **Pandas** : Pour la manipulation des données.
- **NumPy** : Pour les opérations numériques.
- **Scikit-learn** : Pour le prétraitement des données et les algorithmes de machine learning.
- **TensorFlow et Keras** : Pour l'entraînement des modèles impliqués dans l'approche proposée.

En complément à notre environnement d'implémentation, nous avons utilisé Google Colab, une plateforme offerte gratuitement par Google, qui permet d'écrire et d'exécuter des codes en Python via un navigateur. Basé sur Jupyter Notebook, Colab est spécialement conçu pour la formation et la recherche en apprentissage automatique, facilitant l'entraînement de modèles directement dans le cloud. Colab se distingue par son accès gratuit à des processeurs graphiques GPU, offrant ainsi une solution efficace pour les tâches gourmandes en calcul [29].

En outre, notre environnement local pour le développement inclut un ordinateur HP équipé d'un processeur Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz (jusqu'à 1.90 GHz en mode turbo), 8,00 Go de RAM (7,81 Go utilisable), fonctionnant sous Windows 11 Professionnel, système d'exploitation 64 bits.

4.3 Description des ensembles d'entraînement et de test à partir du dataset *Wisconsin* avant et après sa modification

Dans la Phase 01 de notre approche, nous utilisons le dataset Wisconsin tel qu'il existe dans la littérature, avec 32 caractéristiques et 569 cas dont 212 malins et 357 bénins. La colonne "diagnosis" contient des valeurs catégorielles "M" pour malin et "B" pour bénin, mappées à des valeurs

numériques, 0 pour bénin et 1 pour malin. Pour l'entraînement du modèle ANN, les données ont été divisées en deux groupes distincts : le premier est celui correspondant aux cas bénins et le deuxième est celui associé aux cas malins. Pour chaque groupe, 70 % des données sont sélectionnées aléatoirement pour constituer l'ensemble d'entraînement, assurant le maintien des proportions des classes. Les 30 % restants des données sont utilisées pour les tests. De plus, 20 % des 70 % dédiés à l'entraînement sont réservés pour la validation du modèle.

Dans la Phase 02 de notre approche, nous sélectionnons les mêmes 70 % des 212 cas malins que nous avions utilisés pour l'entraînement de l'ANN afin d'entraîner K-means et RF. Le test de ces derniers modèles se fera sur l'ensemble des données prédites durant la Phase 01 correspondant aux cas de tumeurs malignes.

Notre objectif est non seulement de détecter la présence d'une tumeur maligne (cancer), mais aussi de déterminer le stade de la maladie. Pour ce faire, nous avons besoin de données plus détaillées sur les caractéristiques des tumeurs malignes.

4.4 Présentation des métriques utilisées

Pour évaluer la performance de notre approche, nous avons utilisé une méthode d'évaluation pour comparer les résultats de notre modèle avec les *ground-truths*. Un ensemble de quatre métriques, similaire à celles employées par Bishop et al. [17] dans l'évaluation de leur méthode, est utilisé pour cette comparaison. Ces métriques sont : Précision (en anglais *Precision*), Rappel (en anglais *Recall*), Exactitude (en anglais *Accuracy*) et F-mesure (en anglais *F1 Score*). Dans le calcul de ces métriques, on utilise le Vrai Positif, le Faux Positif, le Vrai Négatif et le Faux Négatif définis comme suit :

- **Vrai Positif (True Positive, TP)** : Un vrai positif se produit lorsqu'un modèle prédit correctement la classe positive.
- **Faux Positif (False Positive, FP)** : Un faux positif se produit lorsqu'un modèle prédit incorrectement la classe positive.
- **Vrai Négatif (True Negative, TN)** : Un vrai négatif se produit lorsqu'un modèle prédit correctement la classe négative.
- **Faux Négatif (False Negative, FN)** : Un faux négatif se produit lorsqu'un modèle prédit incorrectement la classe négative.

Les formules des quatre métriques sont données par les équations suivantes [17] :

- **Précision (Precision)** : La précision mesure la proportion des prédictions positives correctes parmi toutes les prédictions positives faites par le modèle.

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4.1}$$

- **Exactitude (Accuracy)** : L'exactitude mesure la proportion des prédictions correctes (po-

sitives et négatives) parmi toutes les prédictions.

$$\text{Exactitude} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}}. \tag{4.2}$$

- **Rappel (Recall) ou Sensibilité (Sensitivity)** : Le rappel mesure la proportion des vrais positifs détectés par le modèle parmi tous les éléments réellement positifs.

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4.3}$$

- **F-mesure (F1 Score)** : La F-mesure est la moyenne harmonique de la précision et du rappel, ce qui permet de considérer à la fois les faux positifs et les faux négatifs dans une seule mesure.

$$\text{F1 Score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}. \tag{4.4}$$

4.5 Évaluation de l'approche proposée

Cette section est dédiée aux tests et à l'évaluation des deux phases de l'approche proposée. En effet, nous testons et évaluons la première phase sur le dataset Wisconsin avant modification, et la deuxième phase avec le dataset Wisconsin après avoir été complété, en considérant les métriques précédemment présentées. La première phase est comparée à des méthodes concurrentes de l'état de l'art, tandis que la deuxième phase, étant donné qu'il n'y a pas d'articles dans la littérature sur la classification des tumeurs malignes en trois stades, est évaluée en comparant les résultats de K-means avec une autre méthode d'apprentissage supervisé, à savoir RF.

4.5.1 Évaluation et comparaison de la première phase (classification bénigne/maligne)

Les résultats des différentes métriques pour les deux articles de l'état de l'art, celui de Assegie et al. [15] et celui de Nguyen et al. [54], ainsi que pour notre modèle ANN après l'évaluation de la première phase de classification bénigne/maligne sont présentés dans le tableau suivant :

Méthodes / Métriques	Précision	Recall	F1 Score	Accuracy
Assegie et al. [15]	/	/	/	94.35%
Nguyen et al. [54]	/	/	/	99.82%
Notre modèle ANN	97.34%	95.31%	96.17%	96.49%

Tableau 4.1 – Résultats de comparaison du modèle ANN de la Phase 01 de notre approche avec deux méthodes de l'état de l'art.

4.5.1.1 Discussion des résultats de la première phase

Dans cette partie, nous discutons des résultats représentés dans le Tableau 4.1, qui présentent les valeurs des différentes métriques pour le modèle ANN de la Phase 01 de notre approche et celles associées au méthodes de Assegie et al. [15] et Nguyen et al. [54]. Notons que nous avons choisi ces deux articles pour la comparaison parce que nous avons utilisé le même dataset *Wisconsin* pour l'évaluation, limitant ainsi la comparaison à la seule métrique accuracy, puisque ces articles n'ont pas mentionné et utilisé les autres métriques.

À partir du Tableau 4.1, nous observons que notre modèle ANN atteint une accuracy de 96.49%, surpassant l'approche de Assegie et al. [15] qui utilise l'algorithme KNN (94.35% d'accuracy) mais étant légèrement inférieure à l'approche de Nguyen et al. [54] qui utilise les RFs (forêts aléatoires) avec 99.82% d'accuracy.

Notre modèle ANN se distingue par l'utilisation de Réseaux de Neurones Artificiels , avec des techniques avancées de normalisation et de régularisation, ce qui permet une flexibilité dans la représentation des caractéristiques complexes des données de cancer du sein. En comparaison, l'approche KNN se concentre sur l'optimisation des hyperparamètres pour améliorer la précision de la détection, tandis que l'approche basée sur les RFs combine la sélection de caractéristiques avec un modèle d'apprentissage en utilisant l'élimination séquentielle vers l'arrière (en anglais *Sequential Backward Elimination*) pour obtenir une précision très élevée.

En plus de l'accuracy, notre modèle ANN présente des valeurs élevées de Précision (97.34%), de Recall (95.31%) et de F1 Score (96.17%). Ces métriques montrent que notre modèle est non seulement précis mais aussi équilibré en termes de précision et de recall, ce qui est crucial pour réduire les erreurs dans les diagnostics cliniques.

Par ailleurs, il est important de noter que, puisque les trois modèles ont été évalués sur un ensemble de données relativement petit, les résultats de ces différentes métriques sont relativement proches. Par conséquent, il est difficile de juger de manière définitive qu'un modèle est supérieur à l'autre. Une évaluation sur un ensemble de données plus large serait nécessaire pour tirer des conclusions plus robustes.

4.5.2 Évaluation et comparaison de la deuxième phase (Classification de la tumeur maligne en trois stades)

Les résultats des différentes métriques pour notre modèle de la Phase 02 de classification par stades de la tumeur maligne (précoce, intermédiaire et avancé), en utilisant *K-means* ainsi que *Random Forest (RF)*, sont présentés dans le tableau suivant :

Modèles / Métriques	Précision	Recall	F1 Score	Accuracy
K-means	95.99%	95.96%	95.92%	93.10%
RF	94.58%	95.38%	94.90%	93.10%

Tableau 4.2 – Résultats de comparaison des modèles K-means et RF de la Phase 02.

4.5.2.1 Discussion des résultats de la deuxième phase

Dans cette partie, nous discutons les résultats représentés dans le Tableau 4.2, correspondant à la comparaison des performances de *K-means* et RF pour la classification par stade de la tumeur maligne. Notons que nous avons utilisé le dataset *Wisconsin* que nous avions complété pour l'évaluation et les mêmes métriques de comparaison.

À partir du Tableau 4.2, nous observons que les résultats montrent que le modèle *K-means* surpasse légèrement le modèle RF sur les métriques de précision, de rappel, et de F1 Score, tandis que les deux modèles affichent une précision similaire. Cependant, étant donné que la différence entre les métriques est très faible, il est difficile de conclure que K-means est véritablement supérieur à RF sur ce dataset spécifique.

Ces résultats suggèrent que, bien que K-means ait montré une légère supériorité dans cette analyse, cette différence pourrait être due à la petite taille du dataset, ce qui pourrait rendre les résultats sensibles aux variations aléatoires. De plus, le modèle RF, qui bénéficie de l'information étiquetée, pourrait montrer des avantages plus nets dans des scénarios avec des datasets plus grands ou plus complexes. Il est donc crucial de prendre en compte la taille et la nature du dataset lors de l'évaluation des performances des modèles.

4.6 Conclusion

Dans ce chapitre, nous avons évalué notre approche de classification des tumeurs du sein en deux étapes correspondant à Phases 01 et 02 : d'abord entre bénignes et malignes, puis en classant les tumeurs malignes en trois stades (précoce, intermédiaire, ou avancé). Pour la Phases 01, notre modèle ANN a montré une performance compétitive avec une précision de 96.49%, surpassant l'approche KNN et étant proche de celle utilisant RF. Pour la Phases 02, en termes de classification par stades des tumeurs malignes, RF a légèrement surpassé K-means, obtenant un meilleur équilibre entre précision et rappel, avec un F1 Score supérieur. Cependant, l'accuracy était identique pour les deux modèles à 93.10%. Ces résultats indiquent que, bien que RF soit plus performant dans cette tâche spécifique, l'évaluation sur un petit ensemble de données limite la généralisation. Ainsi, une évaluation sur un ensemble plus large serait nécessaire pour confirmer ces conclusions. Notre approche démontre une bonne robustesse et flexibilité, ouvrant des perspectives pour des améliorations futures et des applications cliniques.

Conclusion générale

La détection et la classification du cancer du sein représentent des enjeux cruciaux pour un diagnostic précis et un traitement efficace. Dans le cadre de ce mémoire, nous avons cherché à innover et à raffiner des méthodes existantes en combinant des techniques avancées de deep learning, machine learning et de statistiques.

Après avoir étudié et analysé les notions de base liées à la problématique traitée, nous avons effectué un état de l'art des travaux existants dans la littérature, en les classant et en les analysant selon une classification que nous avons proposée. Cette démarche nous a permis de mieux comprendre les avantages et les limites des approches clés proposées dans la littérature, tout en identifiant des perspectives d'amélioration pour ce domaine.

Nos contributions se sont articulées autour de deux axes principaux. Premièrement, nous avons appliqué un Réseau de Neurones Artificiels (ANN) pour classifier les tumeurs du sein en bénignes ou malignes. Cette approche, largement répandue dans la littérature, a été choisie pour sa capacité à extraire des caractéristiques complexes et pertinentes à partir de données médicales, améliorant ainsi la précision du diagnostic.

Pour affiner notre méthode, nous avons introduit une seconde phase de classification utilisant l'algorithme *RF/K-means* pour classifier les tumeurs malignes en trois stades distincts : précoce, intermédiaire et avancé.

Les résultats obtenus démontrent que l'intégration des ANN et de *RF/K-means* améliore significativement la précision de la classification des tumeurs. Cette approche combinée permet non seulement de déterminer la nature maligne ou bénigne des tumeurs, mais aussi d'évaluer avec précision le stade de progression des tumeurs malignes. Cette évaluation détaillée est essentielle pour personnaliser les traitements en fonction du stade spécifique de la maladie.

Enfin, ce mémoire met en lumière l'importance d'une approche intégrative utilisant les techniques de deep learning, machine learning et statistiques pour améliorer la classification du cancer du sein. Les résultats prometteurs obtenus suggèrent que cette méthode pourrait conduire à des diagnostics plus précis et à des traitements plus efficaces, contribuant ainsi à une meilleure prise en charge clinique des patients.

A l'issue de ce travail, nous avons dégagé plusieurs perspectives qui sont intéressantes à considérer dans le cadre des travaux futurs pour essayer d'améliorer notre approche. Dans ce

sens, nous citons ci-dessous quelques unes :

- Améliorer l'approche proposée par la combinaison d'autres méthodes d'apprentissage et de faire appel au technique d'ensemble learning.
- Tester l'approche proposée sur un dataset de plus grande taille et dont la colonne ajoutée "Stage" soit validée par des experts du domaine.
- Application de l'approche proposée en pratique, notamment pour tirer des conclusions sur ses performances réelles.

Bibliographie

[1] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data, (Consulté le 10 Avril 2024).

[2] https://acteurdemasante.lu/fr/cancer-du-sein/le-sein-a-la-decouverte-de-son-anatomie/, (Consulté le 10 Mars 2024).

[3] https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer, (Consulté le 11 Mars 2024).

[4] https://www.uicc.org/news/globocan-2020-global-cancer-data, (Consulté le 11 Mars 2024).

[5] https://cancer.ca/fr/cancer-information/cancer-types/breast/treatment/surgery, (Consulté le 11 Mars 2024).

[6] https://www.lemagit.fr/definition/Apprentissage-non-supervise, (Consulté le 16 Mars 2024).

[7] https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501309-apprentissage-non-supervise/, (Consulté le 16 Mars 2024).

[8] https://pathology.jhu.edu/breast/staging-grade/, (Consulté le 23 Mars 2024).

[9] https://www.ottawahospital.on.ca/wp-content/uploads/2020/09/MRI_slide4.1.jpg, (Consulté le 23 Mars 2024).

[10] https://ledigitalizeur.fr/ia/lapprentissage-par-renforcement/, (Consulté le 23 Mars 2024).

[11] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html, (Consulté le 25 juin 2024).

[12] H. Abdi and L.J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews : Computational Statistics*, 2(4) :433–459, 2010.

[13] A. Alarabeyyat, M. Alhanahnah, et al. Breast cancer detection using k-nearest neighbor machine learning algorithm. In *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*, pages 35–39. IEEE, 2016.

[14] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6) :e0177544, 2017.

[15] T. Assegie and T. Admassu. An optimized k-nearest neighbor based breast cancer detection. *Journal of Robotics and Control (JRC)*, 2(3) :115–118, 2021.

[16] H. Attoumi. Approche à base d'alliances dans les graphes pour la segmentation d'images : Application pour l'extraction de tumeurs dans des images irm du sein. Master's thesis, Université A/Mira de Béjaia, Béjaia, Algérie, septembre 2022.

[17] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, USA, 2006. Assistant Director, Microsoft Research Ltd, Cambridge, CB3 0FB, U.K.

[18] F. Bouchebbah. *A new levels propagation approach to image segmentation : theory and its application to 2D/3D breast MR images*. Phd thesis, University of Bejaia, Bejaia, Algeria, July 2020.

[19] L. Breiman. Random forests. *Machine Learning*, 45 :5–32, 2001.

[20] S.R. Chery and M. Dahlbom. *PET : Physics, Instrumentation, and Scanners*. Springer, New York, 1st edition, 2006.

[21] B. Courbière and X. Carcopino. *Stratégie thérapeutique - Cancer du sein*, pages 775–785. Vernazobers-Grego, 99 bd de l'Hôpital, 75013 Paris, ikb edn edition, 2022.

[22] B. Courbière and X. Carcopino. *Stratégie thérapeutique - Cancer du sein*, page 776. Vernazobers-Grego, 99 bd de l'Hôpital, 75013 Paris, ikb edn edition, 2022.

[23] B. Courbière and X. Carcopino. *Stratégie thérapeutique - Cancer du sein*, page 777. Vernazobers-Grego, 99 bd de l'Hôpital, 75013 Paris, ikb edn edition, 2022.

[24] B. Courbière and X. Carcopino. *Stratégie thérapeutique - Cancer du sein*, pages 779–781. Vernazobers-Grego, 99 bd de l'Hôpital, 75013 Paris, ikb edn edition, 2022.

[25] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1) :21–27, 1967.

[26] K. Dahmane. *Analyse d'images par méthode de Deep Learning appliquée au contexte routier en conditions météorologiques dégradées*. Thèse de doctorat en vision par ordinateur et reconnaissance de formes, Université Clermont Auvergne, novembre 2020.

[27] B. Dai, R.C. Chen, S.Z. Zhu, and W.W. Zhang. Using random forest algorithm for breast cancer diagnosis. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 449–452. IEEE, 2018.

[28] R. Daoudi-Dabladji. *Classification du cancer du sein par des approches basées sur les systèmes immunitaires artificiels*. PhD thesis, Université Paris-Saclay ; Université d'Evry-Val-d'Essonne, 2016.

[29] M.A.E. Djaballah. Système de prédiction de la consommation d'énergie basé sur le deep learning. Master's thesis, Université de 8 Mai 1945 – Guelma, Faculté des Mathématiques, d'Informatique et des Sciences de la matière, Département d'Informatique, 2021.

[30] N.V. Vapnik E.B. Boser, M.I. Guyon. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, July 1992.

[31] D. Ernst and A. Louette. Introduction to reinforcement learning. 2024.

[32] J. Esther, N. Umadevi, D. Hien, and P. Marc. A novel hybrid k-means and gmm machine learning model for breast cancer detection. *IEEE Access*, 9 :146153–146162, 2021.

[33] A. Gompel. Glande mammaire (pathologies bénignes et malignes). In B. Raccah-Tebeka and G. Plu-Bureau, editors, *La Ménopause en Pratique*, pages 43–49. Elsevier Masson, 2019.

[34] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[36] Y.A. Hamad, K. Simonov, and M.B. Naeem. Breast cancer detection and classification using artificial neural networks. In *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, pages 51–57. IEEE, 2018.

[37] H. Hartmann. La classification et les différents stades du cancer du sein. *Elsan Institut de Radiothérapie et de Radiochirurgie*, 2023.

[38] S. Haykin. *Neural Networks : A Comprehensive Foundation*. Prentice Hall, 1st edition, 1998.

[39] T.G. Herman. *Fundamentals of Computerized Tomography : Image Reconstruction from Projections*. Springer Science and Business Media, London, 2nd edition, 2009.

[40] C. Hsu, S. Kuger, E. Cörek, B. Polat, U. Kämmerer, M. Flentje, and C.S. Djuzenova. *Breast Cancer : Basic and Clinical Research 2014 :8*. Springer, 2015.

[41] C. Huan-Jung and et P.H. Kuo T.H.S. Li. Breast cancer–detection system using pca, multilayer perceptron, transfer learning, and support vector machine. *IEEE Access*, 8 :204309–204324, 2020.

[42] M.S. ichahial and B.A. Thomas. Applying cuckoo search based algorithm and hybrid based neural classifier for breast cancer detection using ultrasound images. *Evolutionary Intelligence*, pages 1–18, 2019.

[43] N. Jayanthi and G. Wadhwa. Classification of breast cancer detection using k-nearest neighbor algorithm trained with wisconsin dataset. *Annals of the Romanian Society for Cell Biology*, pages 4440–4448, 2021.

[44] J.S. Jeyanathan, A. Shenbagavalli, B. Venkatraman, and M. Menaka. Analysis of breast thermograms in lateral views using texture features. In *TENCON 2018-2018 IEEE Region 10 Conference*, pages 2017–2022. IEEE, 2018.

[45] C.T. Johnsto, K.T. Gribbon, and D.G. Bailey. Implementing image processing algorithms on FPGAs, 2004.

[46] P. Kaushik and R. Ratan. Eu-net : Deep reinforcement learning aided breast tumor segmentation and attention based severity classification using fused ultrasound and mammography images. *Journal of Coastal Life Medicine*, 10 :25–45, 2022.

[47] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv :1509.02971*, 2015.

[48] L. Luo, X. Wang, Y. Lin, X. Ma, A. Tan, R. Chan, V. Vardhanabhuti, W.C.W. Chue, K.T. Cheng, and H. Chen. Deep learning in breast cancer imaging : A decade of progress and future directions. 2023.

[49] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1 :281–297, 1967.

[50] G. Maicas, G. Carneiro, A.P. Bradley, J.C. Nascimento, and I. Reid. Deep reinforcement learning for active breast lesion detection from dce-mri. In *International conference on medical image computing and computer-assisted intervention*, pages 665–673. Springer, 2017.

[51] R.M. Mann, N. Cho, and L. Moy. Breast mri : State of the art. *Radiology*, 292 :520–536, 2019.

[52] R.F. Mansour. A robust deep neural network based breast cancer detection and classification. *International Journal of Computational Intelligence and Applications*, 19(01) :2050007, 2020.

[53] M.G. Marmot, D. Altman, D. Cameron, J. Dewar, S. Thompson, and M. Wilcox. The benefits and harms of breast cancer screening : an independent review. *British Journal of Cancer*, 108 :2205–2240, 2013.

[54] C. Nguyen, Y. Wang, and H.N. Nguyen. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. 2013.

[55] D.E. Pisano and J.M. Yaffe. Digital mammography. *Radiology*, 234 :353–362, 2005.

[56] S. Rahangdale, P. Keijzer, and P. Kruit. MBSEM image acquisition and image processing in LabVIEW FPGA. In *International Conference on Systems, Signals and Image Processing, IEEE*, pages 1–4, 2016.

[57] Y.C.A.P. Reddy, P. Viswanath, and B.E. Reddy. Semi-supervised learning : A brief review. *Int. J. Eng. Technol*, 7(1.8) :81, 2018.

[58] Y. Rejani and S.T. Selvi. Early detection of breast cancer using svm classifier technique. *arXiv preprint arXiv :0912.2314*, 2009.

[59] D. Shrinivas, G. Shantala, V. Nitin, G. Puneet, and R. Sharan. Breast cancer detection using gan for limited labeled dataset. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 34–39. IEEE, 2020.

[60] S.S. Thamarai and R. Malmathanraj. Mammogram tumour classification using q learning. *International Journal of Biomedical Engineering and Technology*, 7(4) :339–352, 2011.

[61] V. Vishrutha and M. Ravishankar. Early detection and classification of breast cancer. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing : Theory and Applications (FICTA) 2014 : Volume 1*, pages 413–419. Springer, 2015.

[62] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4) :279–292, 1992.

[63] H. Young, R. Baum, U. Cremerius, K. Herholz, O. Hoekstra, A.A. Lammertsma, J. Pruim, and P. Price. Measurement of clinical and subclinical tumour response using [18f]-fluorodeoxyglucose and positron emission tomography : Review and 1999 eortc recommendations. *European Journal of Cancer*, 35 :1773–1782, 1999.

RÉSUMÉ

Le cancer du sein, principal cancer chez la femme, nécessite une détection précoce et une connaissance de son stade d'avancement pour un traitement efficace. Ainsi, pour une tumeur du sein détectée, il est important d'étudier sa classification pour prodiguer des soins ciblés. Dans ce mémoire, nous avons investigué cette problématique et réalisé plusieurs contributions. La première contribution concerne la réalisation d'un état de l'art sur la détection/classification des tumeurs du sein en suivant une classification que nous avons proposée. La deuxième contribution est dédiée à la proposition d'une méthode de complément du dataset Wisconsin pour les besoins de nos expérimentations. Enfin, la troisième contribution est consacrée à la proposition d'une nouvelle approche de classification des tumeurs en trois stades (précoce, intermédiaire, avancé). Cette approche, qui est constituée de deux phases, combine la méthode d'apprentissage profond ANN, en Phase 01, avec une méthode d'apprentissage automatique (RF ou K-means), en Phase 02. L'approche proposée a été testée et évaluée par phase sur le dataset Wisconsin complété en réalisant des comparaisons appropriées. Les résultats obtenus, en considérant plusieurs métriques de comparaison, ont été satisfaisants.

Mots clés : Classification du cancer du sein, Réseaux de Neurones Artificiels (ANN) ; K-means ; Forêt Aléatoire ; Centiles, Dataset Wisconsin, Diagnostique du cancer du sein.

ABSTRACT

Breast cancer, the main cancer in women, requires early detection and knowledge of its stage for effective treatment. Thus, for a detected breast tumor, it is important to study its classification to provide targeted care. In this document, we have investigated this issue and made several contributions. The first contribution concerns the production of a state-of-the-art review on the detection/classification of breast tumors following a classification that we have proposed. The second contribution is dedicated to proposing a method to complement the Wisconsin dataset for the purposes of our experiments. Finally, the third contribution is devoted to the proposal of a new approach for classifying tumors into three stages (early, intermediate, advanced). This approach, which consists of two phases, combines the ANN deep learning method, in Phase 01, with a machine learning method (RF or K-means), in Phase 02. The proposed approach has been tested and evaluated in phases on the completed Wisconsin dataset by making appropriate comparisons. The results obtained, considering several comparison metrics, were satisfactory.

Key words : Breast cancer classification, Artificial Neural Networks (ANN) ; K-means ; Random Forest ; Percentiles, Wisconsin Dataset, Breast Cancer Diagnosis.