الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
جامعة عبد الحميد بن باديس – مستغانم
Abdel Hamid Ibn Badis University – Mostaganem
كلية العلوم والتكنولوجيا
Faculty of Sciences and Technology
قسم الهندسة الكهربائية
Department of Electrical Engineering

N° d'ordre : M2……/GE/2025

MEMOIRE DE FIN D'ETUDES DE MASTER ACADEMIQUE

Filière : Automatique

Spécialité : **Automatique et Informatique Industrielle**

Thème

Étude comparative de la classification des tumeurs mammaires de la base DDSM par KNN et réseaux de neurones artificiels

Présenté par :

1- **TOUATIA HAMMOU**

Soutenu le 30/ 09 / 2025 devant le jury composé de :

Présidente :	Mme. A. BENCHELLEL	Maitre de Conférences "B"	Université de Mostaganem
Examinateur :	Mr. M. REBHI	Maître Assistant "A"	Université de Mostaganem
Examinateur :	Mr. M. BENAOUALI	Maître Assistant "A"	Université de Mostaganem
Encadrante :	Mme. K. BERRADJA	Maitre de Conférences "A"	Université de Mostaganem

Année universitaire 2024 / 2025

Dédicaces

Tout d'abord, je tiens à remercier الله pour m'avoir donné la force et le courage de mener bien ce modeste travail.

A mon père Menouare et ma mère Djouhar pour avoir été à mes côtés tout au long de mon parcours académique et pour m'avoir encouragé dans ce long cheminement.

A mes très chères sœurs, Hanane, Imene, Fatima Zohra, ceux *qui m'ont soutenues, réconforté et sont restés à mes côtés malgré tous les obstacles tout au long de mes études*

A toute la famille HAMMOU

A tous mes professeurs qui ont contribué à la réalisation de cet humble travail

A tous les étudiants de la promotion Automatique 2025.

TOUATIA HAMMOU

Remerciements

*J'*adresse mes remerciements à mon encadrante, Dr. KHADIDJA BERRADJA pour son aide précieuse et ses conseils qui m'ont été d'une grande utilité dans la réalisation de ce mémoire.

*J'*exprime tout mon respect et ma gratitude aux membres du jury qui consacrent leur temps et leur attention à l'évaluation de ce travail. Je remercie tout particulièrement Mme A. BENCHELLEL pour avoir accepté de présider le jury, Mr. M. REBHI ainsi que Mr. M. BENAOUALI pour avoir accepté d'examiner ce mémoire.

*J*e tiens enfin à remercier l'ensemble de l'équipe pédagogique du département de Génie Electrique pour leur encadrement et leur disponibilité tout au long de mon parcours universitaire.

Table des Matières

Liste des Figures

CHAPITRE I

CHAPITRE II

CHPITRE III

CHPITRE IV

Liste des Tableaux

Liste des Abréviations

ML	Machine Learning
ANN	Artificiels Neurone Network
KNN	k-Nearest Neighbors
CAD	Computer Aided Diagnosis
DDSM	Digital Data Base for screening Mammography
CBIS-DDSM	Curated Breast Imaging Subset of DDSM

Résumé

Le cancer du sein est une cause majeure de mortalité féminine, touchant une femme sur huit. La détection précoce, cruciale pour le pronostic, est améliorée grâce à la mammographie numérique et aux systèmes d'aide au diagnostic assistés par ordinateur (CADx). Cependant, repérer les premières lésions reste difficile, même pour les radiologues. C'est dans ce contexte que ce projet se concentre sur la classification des masses mammaires (lésions bénignes vs malignes) par apprentissage automatique, en utilisant le K-Nearest Neighbors (KNN) et les réseaux de neurones artificiels (ANN). L'objectif poursuivi est de réduire les erreurs de diagnostic, d'apporter un appui décisionnel aux radiologues et d'optimiser la prise en charge des patientes. À cette fin, la comparaison entre les algorithmes KNN et ANN repose sur plusieurs critères de performance : exactitude, précision, sensibilité et spécificité. L'algorithme KNN, bien que simple à mettre en œuvre, demeure sensible au bruit, tandis que les réseaux de neurones artificiels tirent parti de l'apprentissage automatique pour identifier des relations et des structures complexes. Ainsi, le choix de l'algorithme le plus adapté dépendra de son efficacité, conformément aux exigences d'un usage clinique.

Mots clés : cancer du sein, systèmes d'aide au diagnostic, Apprentissage automatique, classification, réseaux de neurones artificiels (ANN), k-plus proches voisins (KNN).

Abstract

Breast cancer is a major cause of female mortality, affecting one in eight women. Early detection, which is crucial for prognosis, has been enhanced through digital mammography and computer-aided diagnostic (CADx) systems. However, identifying early lesions remains challenging, even for radiologists. In this context, the present project focuses on the classification of breast masses (benig vs malignant lesions) using machine learning, specifically K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN). The main objective is to reduce diagnostic errors, provide decision support to radiologists, and optimize patient care. To this end, the comparison between KNN and ANN is based on several performance criteria: accuracy, precision, sensitivity, and specificity. While KNN is simple to implement, it remains sensitive to noise, whereas Artificial Neural Networks leverage machine learning to capture complex and non-linear relationships. Therefore, the choice of the most suitable algorithm depends on its effectiveness, in line with the requirements of clinical use.

Keywords : Breast cancer, diagnostic support systems, machine learning, classification, artificial neural networks (ANN), k-nearest neighbors (KNN).

ملخص

السرطان الثدي يُعدُّ سبباً رئيسياً في وفيات النساء، حيث يصيب امرأة من بين كل ثماني نساء. تُعتبر الكشف المبكر - الذي يعدُّ حاسماً في تحديد سير المرض وتوقعاته - أكثر فعالية بفضل التصوير الشعاعي الرقمي للثدي وأنظمة الدعم التشخيصي المعتمدة على الحاسوب (CADx).إلا أن تحديد الآفات الأولى لا يزال يشكل تحدياً كبيراً، حتى لأخصائي الأشعة. في هذا السياق، يركز هذا المشروع على تصنيف الأورام الثدية (الأورام الحميدة مقابل الأورام الخبيثة) باستخدام تقنيات التعلم الآلي، بالاعتماد على خوارزميات الجار الأقرب (KNN)وشبكات العصب الاصطناعية (ANN). الهدف من ذلك هو تقليل أخطاء التشخيص، وتقديم دعم اتخاذ القرار للأخصائي الأشعة، وتحسين رعاية المرضى. لتحقيق هذا، تُجرى مقارنة بين خوارزميتي KNN و ANNاستناداً إلى عدة مقاييس أداء تشمل الدقة، والتحسُّن، والحساسية، والخصوصية. بالرغم من بساطة تطبيق خوارزميةKNN ، إلا أنها تبقى حساسة للضوضاء، في حين تستفيد شبكات العصب الاصطناعية من التعلم الآلي للكشف عن العلاقات والبنى المعقدة. وبناءً عليه، يعتمد اختيار الخوارزمية الأنسب على مدى فاعليتها، تماشياً مع متطلبات الاستخدام السريري.

الكلمات المفتاحية: سرطان الثدي، أنظمة دعم التشخيص، التعلم الآلي، التصنيف، شبكات العصب الاصطناعية(ANN) ، الجار الأقرب (KNN) .

Introduction Générale

Le cancer du sein représente le cancer le plus répandu chez les femmes à l'échelle mondiale, mais également le deuxième cancer le plus courant tous sexes confondus. Un diagnostic précoce et précis joue un rôle crucial dans l'amélioration du pronostic et de la survie des patientes. Dans ce contexte, les technologies d'imagerie médicale, en particulier la mammographie, constituent un outil de dépistage fondamental. Toutefois, l'interprétation des images reste une tâche complexe, sujette à des erreurs humaines et nécessitant une expertise spécialisée. Face à ces défis, l'intégration de l'intelligence artificielle, et plus spécifiquement des méthodes d'apprentissage automatique (machine learning), ouvre de nouvelles perspectives prometteuses dans l'automatisation et l'optimisation du diagnostic. Le machine learning permet aux systèmes informatiques d'apprendre à partir de données médicales historiques afin de détecter, classifier et prédire la nature des anomalies observées sur les images mammographiques, sans avoir été explicitement programmés pour chaque cas particulier.

Parmi les nombreuses techniques disponibles, certaines méthodes supervisées, telles que les réseaux de neurones artificiels (ANN) et l'algorithme des k plus proches voisins (KNN), se distinguent par leur efficacité dans les tâches de classification binaire, comme la distinction entre masses bénignes et malignes. En s'appuyant sur des bases de données d'images médicales annotées, ces approches peuvent être entraînées à reconnaître des schémas caractéristiques du cancer du sein et à assister les professionnels de santé dans leur prise de décision.

Ce travail s'inscrit dans cette dynamique en explorant l'application de ces deux techniques de classification sur la base de données CBIS-DDSM, dans le but d'évaluer leurs performances respectives à travers des critères tels que la précision, la sensibilité ou encore la spécificité. Il s'agit ainsi de contribuer à l'amélioration des outils d'aide au diagnostic en cancérologie mammaire grâce à l'intelligence artificielle.

Ce mémoire s'articule autour de quatre chapitres qui présentent les différents aspects de ce travail.

Le premier chapitre est dédié à la présentation des notions fondamentales liées à l'anatomie du sein, ainsi qu'à la description de ses principales structures. Il propose également un aperçu des masses mammaires, en abordant leurs caractéristiques, leur classification, ainsi qu'une

introduction à la mammographie radiologique, qui constitue la méthode d'imagerie de référence pour le dépistage du cancer du sein.

Le deuxième chapitre, intitulé « Machine Learning », présente les principes fondamentaux de l'apprentissage automatique, un domaine central de l'intelligence artificielle. Il y est question de la manière dont les machines peuvent améliorer leurs performances à partir de données, sans être explicitement programmées pour chaque tâche. Ce chapitre retrace également l'évolution historique du Machine Learning, en explorant les différentes approches d'apprentissage, les principaux défis rencontrés, ainsi que ses domaines d'application, mettant ainsi en lumière une technologie en pleine expansion dans de nombreux secteurs.

Le troisième chapitre est consacré à l'étude des méthodes de classification utilisées dans ce travail, à savoir l'algorithme des k plus proches voisins (KNN) et le réseau de neurones artificiels (ANN). Après une présentation des principes théoriques de chaque méthode, leurs mécanismes de fonctionnement seront détaillés, en mettant en évidence leurs avantages, leurs limites, ainsi que les contextes dans lesquels elles sont les plus efficaces. Ce chapitre constitue ainsi une base essentielle pour comprendre leur mise en œuvre dans le cadre de la classification du cancer du sein.

Le dernier chapitre, présente l'application pratique des deux méthodes de classification : les réseaux de neurones artificiels (ANN) et l'algorithme des k plus proches voisins (KNN) sur la base d'images mammographiques annotées DDSM. Ce chapitre décrit comment ces techniques ont été utilisées pour distinguer les masses mammaires bénignes des malignes, en évaluant leur capacité à identifier et à caractériser correctement les anomalies. Les performances des deux approches sont ensuite comparées à l'aide de plusieurs indicateurs d'évaluation (tels que la précision, la sensibilité, la spécificité, etc.), afin de déterminer laquelle offre les meilleurs résultats dans le contexte du diagnostic assisté du cancer du sein.

Enfin, une conclusion et les perspectives envisagées sont présentées en dernier lieu.

CHAPITRE : I

GÉNÉRALITÉS SUR LE CANCER DU SEIN

I.1 Introduction

L'objectif de ce chapitre est d'établir le cadre médical de notre étude. Pour cela, nous commencerons par une description de l'architecture anatomique de la glande mammaire, afin de mieux comprendre les origines possibles des anomalies observées. Ensuite, nous mettrons en évidence la technique d'imagerie de référence dans le dépistage du cancer du sein : la mammographie radiologique.

L'étude des pathologies mammaires les plus courantes est structurée autour des deux grandes catégories de signes radiologiques : les masses et les calcifications. Une attention particulière sera portée à l'analyse des masses mammaires, qui représentent un enjeu central dans le diagnostic différentiel entre lésions bénignes et malignes. Ce focus permettra de mieux situer le contexte médical de notre application.

I.2 Anatomie et pathologies du sein

Chez la femme, le sein est un organe destiné à produire du lait pour nourrir l'enfant, qui se développe à la puberté sous l'influence des hormones. Chez l'homme, le sein est un reliquat d'organe mammaire, semblable au sein féminin prépubère, qui présente une morphologie relativement similaire à celle des femmes, mais dans une dimension toutefois atrophiée. Aussi, l'homme peut être sujet à certaines pathologies du sein comme la femme bien que cela soit moins courant. [1]

I.2.1 Définition du sein

Le sein a une structure complexe : il est composé de 15 à 20 compartiments, séparés par du tissu adipeux, qui lui donnent la forme qu'on lui connaît. Chacun de ces compartiments est constitué de canaux et de lobules [2] (voir la figure I.1).

Pour mieux visualiser cette structure, on peut imaginer un arbre avec plusieurs branches (les canaux) rattachées à un point central (le mamelon). Les lobules se situent à l'extrémité de ces branches. A la puberté, la jeune fille observe des changements dans la forme et le volume de ses seins sous l'action des hormones sexuelles. Celles-ci ont une influence sur les seins tout au long de la vie. Deux types d'hormones jouent un rôle important et ont une action complémentaire : les œstrogènes et la progestérone.[3]

1. Clavicule
2. Muscle
3. Tissu conjonctif
4. Tissu adipeux
5. Aréole
6. Mamelon
7. Canaux
8. Glandes (lobules)
9. Peau
10. Côtes

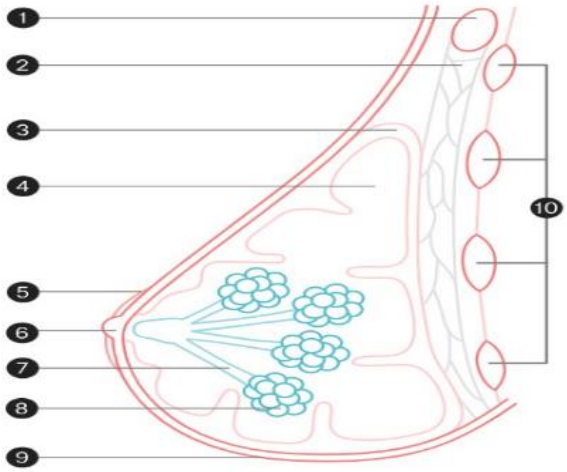


Figure I.1 : Éléments importants dans la composition du sein [4].

I.2.2. Structure du sein et tissus avoisinants

Le sein est situé au niveau de la cage thoracique, sur le muscle grand pectoral qui s'étend de la clavicule au sternum, et est lié au creux axillaire. D'un point de vue médical, les seins sont divisés en quatre zones appelées quadrants définies par des lignes invisibles qui n'ont pas de « frontières » anatomiques, mais permettent de localiser aisément d'éventuelles anomalies. Le schéma des quadrants s'établit en plaçant deux lignes perpendiculaires sur le sein qui s'étendent de haut en bas et de gauche à droite, et se croisent au niveau du mamelon [5].

Le quadrant situé en haut du côté de l'aisselle est le supéro-externe, celui en haut du côté du sternum est le supéro-interne, le quadrant en bas du côté de l'aisselle est l'inféro-externe et celui en bas du côté du sternum est l'inféro-interne. Les schémas des quadrants sont symétriques, aussi le quadrant supéro-externe, par exemple, se situe en haut à droite pour le sein droit, et en haut à gauche pour le sein gauche. Si cette délimitation est importante à connaître, c'est notamment parce que le quadrant supéro-externe est le siège de prédilection des cancers du sein, une zone qui est donc particulièrement à surveiller lors des examens cliniques et de l'autopalpation. À noter que les cancers peuvent se développer à cheval sur plusieurs quadrants, et que les cancers de la région centrale (proche du mamelon) sont les seconds plus répandus après ceux du quadrant supéro-externe [6].

I.2.3. Principaux éléments de la glande mammaire

Le sein, ou glande mammaire, est composé de différents éléments qui possèdent chacun une fonction propre et peuvent être sujets à diverses pathologies. Le sein est en grande partie composé de tissus adipeux qui n'impactent pas la production de lait, mais influencent hautement le volume du sein. Les alvéoles sont de petits sachets microscopiques produisant le lait maternel et regroupés dans des glandes nommées lobules [7].

Les canaux galactophores sont destinés à faire circuler le lait depuis les alvéoles vers les mamelons. Le mamelon est composé de tissus musculaires permettant de véhiculer le lait. L'aréole est la zone pigmentée qui entoure le mamelon et contient des glandes aréolaires sécrétant un liquide protecteur pour faciliter la lactation. Les ligaments de Cooper forment le tissu conjonctif qui entoure et soutient le sein pour lui permettre de garder sa structure. Enfin, le sein comprend un réseau de vaisseaux sanguins permettant sa vascularisation, ainsi que des vaisseaux lymphatiques jouant un rôle essentiel dans la santé de l'organe mammaire (voir la figure I.2).

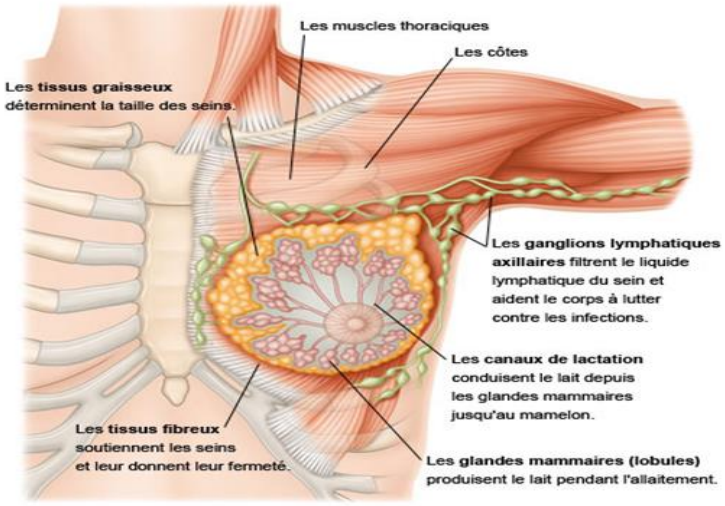


Figure I.2 : Principaux éléments de la glande mammaire [8].

I.2.4. Les principales maladies du sein

I.2.4.1. Pathologies bénignes du sein

Le sein peut être sujet à de multiples affections bénignes, dont l'étendue et l'aspect varient chez les différentes patientes, et qui dégénèrent rarement en lésions cancéreuses. Les kystes mammaires, lipomes et tumeurs bénignes peuvent provoquer l'apparition de masses plus ou moins importantes, accompagnées ou non de douleurs, qui peuvent inquiéter la patiente, mais

sont généralement sans danger. Parmi les tumeurs bénignes, on trouve notamment des fibroadénomes, papillomes, phyllodes, tumeurs à cellules granuleuses, hémangiome… [9].

Des affections d'origines bactériennes peuvent aussi intervenir, provoquant des mastites nécessitant un traitement médical [10].

I.2.4.2. Un mamelon surnuméraire

Les mamelons surnuméraires, ou accessoires, sont une anomalie fréquente mais méconnue. Bien qu'ils ne présentent aucun danger pour la santé, ils peuvent causer une gêne esthétique ou physique. Leur correction repose sur des interventions chirurgicales comme l'ablation, la réduction du tissu mammaire ou le lipofilling, permettant d'améliorer la symétrie et le confort. L'opération, rapide et peu invasive, aide à restaurer l'harmonie des seins et à renforcer l'estime de soi. Après la chirurgie, des soins post-opératoires sont nécessaires pour une bonne cicatrisation et des résultats optimaux. Cette correction apporte un bien-être physique et psychologique durable [11].

I.2.4.3. Pathologies cancéreuses du sein

Le carcinome canalaire est le cancer du sein le plus courant. Il touche les canaux mammaires et peut prendre deux formes différentes en fonction de son agressivité (in situ ou infiltrant). Plus rare, le cancer inflammatoire est une forme de cancer du sein qui affecte le système lymphatique [11].

Le carcinome lobulaire touche les lobules du sein, et peut également adopter une forme in situ ou infiltrante. La maladie de Paget est une autre forme rare de cancer du sein qui se manifeste à travers des symptômes cutanés et une altération de l'aspect du sein. Le sein peut également être atteint de lymphomes diffus, sarcomes des tissus mous, carcinomes adénoïdes kystiques, carcinosarcomes et autres types de tumeurs cancéreuses rares [12].

I.2.5. Définition d'une tumeur

Le mot "tumeur" est un terme générique, correspondant au développement d'un tissu nouvellement formé au sein d'un tissu normal. La tumeur est causée par une anomalie du développement cellulaire. Il existe deux principales classes de tumeurs : les tumeurs bénignes et les tumeurs malignes ou cancers [13].

Contrairement aux idées reçues, le terme de tumeur est utilisé pour des pathologies cancéreuses, mais également pour d'autres productions à caractère bénin. Une tumeur maligne,

est un amas de cellules cancéreuses. Une tumeur bénigne n'est pas un cancer, contrairement à une tumeur maligne. Elle se développe lentement, localement, sans produire de métastases, et ne récidive pas si elle est enlevée complètement. Un kyste peut être considéré comme une tumeur bénigne. (Voir le Tableau I.1**)** ci-dessous exprime les critères de distinction entre les tumeurs bénignes et malignes [14] :

Tumeur bénigne	Tumeurs maligne
Bien limitée.	Mal limitée.
Encapsulée.	Non encapsulée.
Histologiquement semblable au tissu d'origine.	Plus ou moins semblable au tissu d'origine.
Cellules régulières.	Cellules irrégulières (cellules cancéreuses).
Croissance lente.	Croissance rapide.
Refoulement sans destruction des tissus voisins.	Envahissement des tissus voisins.
Pas de récidive locale après exérèse complète.	Récidive possible après exérèse supposée totale.
Pas de métastase.	Métastase(s).

Tableau I.1 : Critère de distinction entre tumeurs bénignes / malignes [15].

I.3. Cancer du sein

Le cancer du sein est une tumeur maligne qui se développe dans les cellules du sein, provoquant la destruction et l'envahissement des tissus. Cette tumeur peut potentiellement se propager aux tissus voisins, et parfois envahir d'autres parties du corps humain : on parle alors de métastases. Une altération du génome des cellules cancéreuses entraîne un comportement anarchique de la cellule avec une prolifération non maîtrisée de la croissance cellulaire. Cette prolifération cellulaire ne respecte plus les codes avec une infiltration des tissus et des vaisseaux environnants. Nous mettrons à part, les lésions mammaires dites in situ qui ont un comportement localement agressif, mais avec un risque tardif de maladie générale. Concept de maladie locale vs maladie générale [16].

Fort heureusement, la grande majorité des pathologies mammaires ne sont pas cancéreuses avec les lésions kystiques mammaires, les tumeurs bénignes tels les fibroadénomes, lésions qui très exceptionnellement peuvent devenir cancéreuses [17].

I.3.1. Les facteurs de risque principaux du cancer du sein

Il y a quatre principaux facteurs de risque de développer un cancer du sein. L'âge est le premier, ce cancer apparaissant souvent autour de 60 ans. Ensuite, il y a les prédispositions génétiques : lorsque plusieurs personnes d'une même famille sont atteintes du même cancer, il peut s'agir d'un cancer héréditaire. C'est le cas de 5 à 10 % des cancers du sein. Dans cette situation, une mutation dite « de prédisposition » est transmise de génération en génération. Les principaux gènes en cause sont les gènes BRCA1 et BRCA2 [18].

Les antécédents familiaux de cancer du sein représentent un troisième facteur de risque. Ils concernent la présence d'au moins un cas isolé dans la famille proche. Enfin, il y a les antécédents personnels de cancer du sein, d'hyperplasies atypiques, de maladies bénignes du tissu mammaire, ou d'exposition à des radiations dans le cadre de précédents traitements [19].

D'autres facteurs de risque plus secondaires ont également été identifiés pour le cancer du sein, comme la puberté précoce et la ménopause tardive, les traitements hormonaux substitutifs de la ménopause, l'absence de grossesse et les grossesses tardives, l'absence d'allaitement, la consommation régulière d'alcool, le tabagisme, la sédentarité et enfin, la surcharge pondérale [20].

I.3.2. Les différents types de cancers du sein

Le terme « cancer du sein » correspond à la présence de cellules anormales dans un sein, cellules qui se multiplient de façon anarchique pour former une tumeur maligne (un « carcinome »). Selon le type de cancer du sein, ces cellules peuvent rester confinées dans le sein ou migrer vers les ganglions avoisinants, voire le reste du corps (métastases).

Selon le type de cellules à l'origine du cancer, et selon l'aspect de la tumeur, on distingue différents cancers du sein. Par exemple, un cancer du sein qui touche les cellules qui bordent les canaux sera dit « canalaire » et un cancer qui affecte les cellules des lobules sera dit « lobulaire » [21].

De plus, certains cancers du sein sont provoqués par des cellules cancéreuses dont la multiplication est favorisée par les hormones sexuelles féminines, les estrogènes. Ce sont les cancers du sein dits « hormonodépendants » ou « hormonosensibles ». En cas de sensibilité des cellules cancéreuses aux estrogènes, le médecin peut bloquer la croissance de la tumeur en supprimant ces hormones par le biais de médicaments adaptés [22].

I.3.2.1. Les cancers du sein non invasifs

On parle de cancer du sein « non invasif » ou « in situ » lorsque la tumeur reste dans le tissu d'origine et n'envahit pas les tissus voisins. Ces cancers sont plus facilement traités que les cancers qui envahissent les tissus voisins (« cancers invasifs », (voir la figure I.3) ci-dessous. Dans 90 % des cas, les cancers du sein non invasifs sont de type canalaire (ils se forment à l'intérieur des canaux de lactation), les 10 % restants étant de type lobulaire [24].

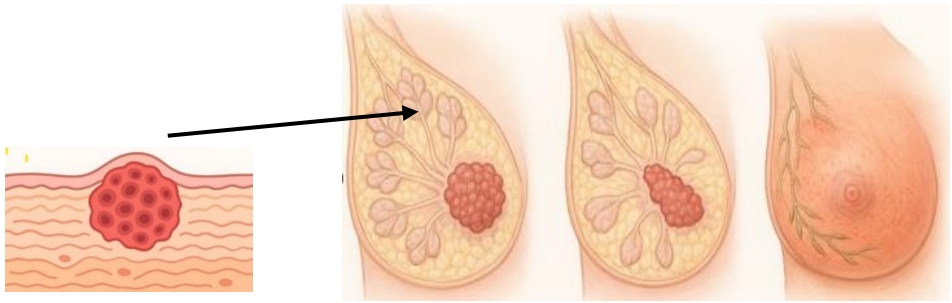


Figure I.3 : Le cancer du sein non invasif [24].

I.3.2.2. Les cancers du sein invasifs

Les cancers du sein sont dits « invasifs » ou « infiltrants » lorsque les cellules cancéreuses ne restent pas confinées à leur lieu d'origine et envahissent les tissus avoisinants, les ganglions locaux (dits « ganglions axillaires » et situés sous l'aisselle) et, parfois, le reste du corps. Ces cancers invasifs touchent essentiellement les canaux, rarement les lobules. Leur pronostic est moins bon que celui des cancers du sein non invasifs [23] (voir la figure I.4).

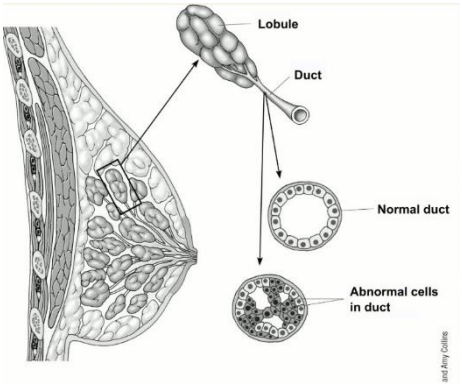


Figure I.4 : Le cancer du sein invasif [24].

I.3.3. Signes et symptômes du cancer du sein

Dans un premier temps, le cancer du sein est asymptomatique chez la plupart des malades ; la détection précoce est donc fondamentale [25].

Il peut néanmoins provoquer une association de différents symptômes, notamment aux stades plus avancés. Les symptômes du cancer du sein sont notamment :

- Une masse ou un épaississement dans le sein, souvent indolore,
- Un changement de la taille, de la forme ou de l'apparence du sein,
- Des fossettes, des rougeurs, une peau d'orange ou d'autres changements cutanés,
- Une modification de l'apparence du mamelon ou de la peau qui l'entoure (aréole),
- Un écoulement mamelonnaire anormal ou sanglant.

En cas de masse anormale dans le sein, même indolore, il convient de consulter un médecin. La plupart des masses de ce type ne sont pas cancéreuses. Le traitement des masses qui sont cancéreuses est plus efficace lorsqu'elles sont de taille réduite et ne se sont pas étendues aux ganglions lymphatiques environnants.

Le cancer du sein peut se propager à d'autres organes et provoquer d'autres symptômes. Le plus souvent, les ganglions lymphatiques situés sous le bras sont le premier site de propagation détectable. Il arrive toutefois qu'on ne sente pas des ganglions lymphatiques porteurs de cancer.

Les cellules cancéreuses peuvent progressivement se propager à d'autres organes comme les poumons et le foie, ainsi qu'au cerveau et aux os. Une fois ces sites atteints, de nouveaux symptômes liés au cancer peuvent apparaître, comme des douleurs osseuses ou des maux de tête.

I.4. Dépistage du cancer du sein

Avant 50 ans, il est essentiel pour les femmes de consulter chaque année un gynécologue pour qu'il procède à un examen clinique des seins. En cas de doute ou d'anomalie, le médecin peut alors programmer des examens complémentaires. Les femmes âgées de 50 à 74 ans sont invitées à se faire dépister du cancer du sein tous les deux ans avec une mammographie, ou si besoin une échographie et ce, sans avance de frais. Plusieurs outils sont utilisés pour établir un diagnostic du cancer du sein [26].

I.4.1. La palpation des seins

La palpation permet la mise en évidence d'une grosseur anormale au niveau des seins et de creux axillaires sous les aisselles. L'autopalpation régulière des seins est également recommandée dans le cadre de la prévention du cancer du sein. En cas d'incertitude, il est essentiel de consulter rapidement un professionnel de santé afin d'obtenir un avis éclairé et, si nécessaire, réaliser des examens complémentaires [27].

I.4.2. L'imagerie médicale

I.4.2.1. L'échographie

L'échographie est un examen utilisant les ultrasons, prescrit lorsque la mammographie a mis en évidence une anomalie, ou lorsque la densité des seins ne permet pas d'avoir une mammographie de qualité. L'imagerie par résonance magnétique (IRM) est quant à elle réalisée pour obtenir des renseignements complémentaires aux informations données par la mammographie et l'échographie [27] (voir la figure I.5).

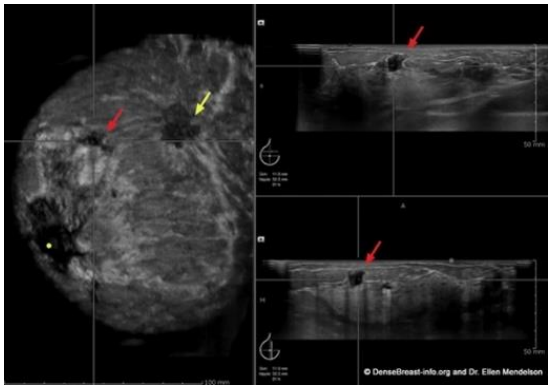


Figure I.5 : Exemple d'échographie du sein [28].

I.4.2.2. La mammographie

La mammographie consiste en une radiographie de chaque sein réalisé avec un appareil appelé « le mammographe » (voir la Figure I.6). Cet examen permet de visualiser les structures internes du sein et donc de détecter toute masse jugée anormale. En pratique, le sein est placé et comprimé entre deux plaques. La pression ressentie peut être inconfortable. Sans conséquence pour la poitrine, ce désagrément s'arrête dès que les plaques sont enlevées. Lors de l'examen, les seins sont soumis à une dose extrêmement faible de rayons X. Deux images ou clichés par sein sont réalisés sous différents angles et analysés par un radiologue [29].

La mammographie ne permet pas toujours de donner d'emblée un diagnostic définitif : elle permet de voir s'il existe une anomalie dans le sein, mais elle ne permet pas de déterminer avec certitude s'il s'agit ou non d'un cancer. La douleur lors de l'examen liée à la compression du sein entre deux plaques est fréquente mais de faible intensité. Le premier risque consécutif à la mammographie de dépistage est le sur-diagnostic. De 20 à 49 % des femmes participantes à un dépistage mammographie régulier auront au moins un résultat faussement positif après 10 examens [29].

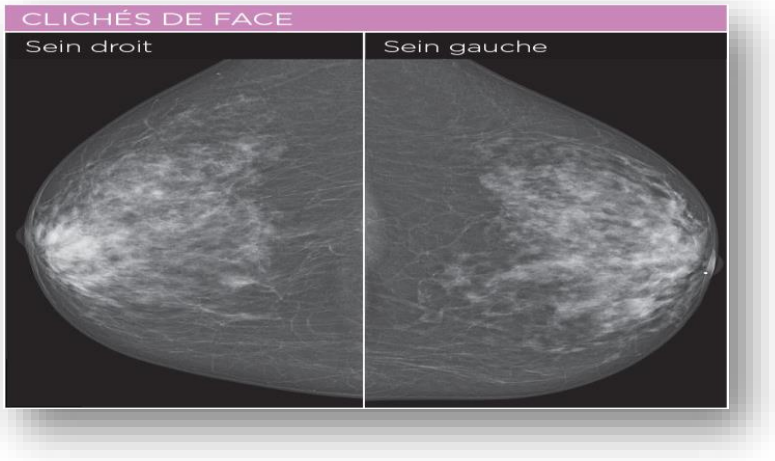


Figure I.6 : Exemple de mammographie du sein [30].

La mammographie numérique est une image électronique produite avant d'être sauvegardé sur ordinateur où elle subira une manipulation du degré de contraste dans les zones les plus denses notamment avant d'être visualisée. On distingue 2 positions de visualisation en mammographie numérique :

• **Lever du bras pour oblique médio latérale (MLO)**

Lors de la visualisation du MLO, le patient est également amené à se pencher vers l'équipement pour une visualisation optimale des tissus (voir la figure I.7). Le bras est tourné à 45 degrés afin de démontrer la quantité maximale de tissu mammaire et de muscle pectoral. Parfois, l'angle est individualisé selon la taille de la poitrine avec une différence de 10 degrés. L'autre sein du patient, non-imageur, est doucement pressé contre le corps et maintenu à l'écart [31].

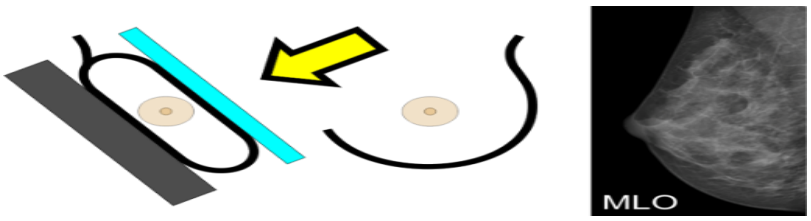


Figure I.7 : Visualisation lever du bras pour oblique médio latérale (MLO) [32].

• **Craniocaudal (CC)**

Pendant la vision CC, le patient est positionné de sorte que le mamelon se trouve approximativement au centre du détecteur (voir la figure I.8). La patiente est obligée de se pencher vers l'appareil pour rapprocher le sein du détecteur. Le bas de la poitrine doit être soutenu et tiré vers le haut afin que les tissus les plus profonds et les plus bas soient inclus dans la vue, du côté en cours d'imagerie, est poussée vers le bas pour relâcher le muscle pectoral de manière à inclure le tissu mammaire dans le quadrant externe. La visualisation du muscle pectoral sur la vue en CC implique qu'aucun tissu le long de la paroi thoracique n'a été exclu [33].

Figure 1.8 : Visualisation craniocaudal (CC) [34].

I.5. Avantages et inconvénients de la mammographie

La mammographie présente à la fois des avantages et des inconvénients.

❖ Avantages

Les avantages pouvant être générés par la mammographie, sont cités ci-dessous [35] :

- ✓ ***La réduction de la mortalité***

La réduction de la mortalité est la raison d'être, d'un programme de dépistage. Les données statistiques après 10 ans de fonctionnement du programme, indiquent une réduction de mortalité de l'ordre de 35 % chez les femmes participant au dépistage du cancer du sein.

- ✓ ***Dépistage précoce***

Un cancer du sein peut être découvert plus tôt grâce au dépistage ce qui est un facteur important pour contribuer à la diminution de la mortalité. Les cancers trouvés au dépistage sont généralement plus petits que ceux trouvés à l'examen physique.

- ✓ ***Chirurgie conservatrice***

Le dépistage précoce du cancer du sein, permet d'avoir une chirurgie moins étendue, permettant de conserver le sein. Il réduit aussi, le risque d'avoir de la chimiothérapie puisque le cancer mis en évidence, étant plus petit.

❖ Inconvénients

Les inconvénients se rapportent à ne pas détecter uniquement, les vrais positifs. Des faux négatifs et positifs, peuvent apparaître lors d'une mammographie [36] :

- ✓ ***Les faux négatifs***

La mammographie n'est pas infaillible. Certains cancers, peuvent passer inaperçus. 10% des signes perçus, sont des faux négatifs.

- ✓ ***Les faux positifs***

La mammographie demandant une investigation supplémentaire qui ne révèlera pas de cancers considérés, comme des faux positifs (10 % des dépistages).

- ✓ **Les résultats faussement positifs ont plusieurs conséquences**

De l'anxiété, une augmentation du nombre d'examens additionnels et requis, ainsi qu'une diminution de la participation des femmes aux dépistages subséquents.

- ✓ ***L'exposition aux rayons x***

Comme toute radiographie, la mammographie expose la patiente, à des rayons X. Ceux-ci, s'ils sont répétés, peuvent conduire à l'apparition d'un cancer, que l'on appelle cancers radio-induits. C'est l'une des raisons pour lesquelles, le dépistage est recommandé uniquement tous les deux ans, à partir de 50 ans si la femme n'a pas de de symptôme ou, de facteurs de risque. Par ailleurs, après 50 ans, la composition des seins se modifie et les doses de rayons nécessaires à la mammographie, sont plus faibles. Cependant Etant utilisés à de très faibles doses lors de la mammographie, l'examen est considéré sans danger.

I.6. Types d'anomalies dans le sein

La généralisation du dépistage du cancer du sein, amène à découvrir beaucoup d'anomalies purement radiologiques. Parmi ces anomalies, on trouve : les calcifications et les masses.

I.6.1. Les calcifications

Les calcifications mammaires sont assez courantes et, la plupart ne sont pas associées au cancer. Afin de s'en assurer, le radiologue étudie leur taille, leur forme et leur disposition à l'aide, d'une mammographie sur laquelle, elles apparaissent souvent sous forme de petits points blancs. Certaines de leurs caractéristiques, comme une forme irrégulière ou certains regroupements, peuvent être suspectes. On cite à cet effet :

- **La macrocalcification mammaire**

La macrocalcification se caractérise par un dépôt dont la taille est supérieure à 1 mm. Ces dépôts grossiers sont plus fréquents chez les femmes de plus de 50 ans et sont généralement bénignes. Ce type de calcification est associé au vieillissement, mais pas uniquement. En effet, elle peut également résulter d'une inflammation ou d'une infection des tissus du sein, ou des lésions causées par un traumatisme ou une chirurgie. La macrocalcification peut également peut aussi être causé par des kystes ou un fibroadénome, une tumeur non cancéreuse du sein [37].

- **La microcalcification mammaire**

Contrairement aux macrocalcifications, les microcalcifications ont une taille inférieure à 1 mm. Ses minuscules dépôts de calcium dans le sein peuvent être bénins ou malins.

Les microcalcifications sont le signe d'une activité accrue de certaines cellules des tissus glandulaires. En s'activant, se développant et se divisant, les cellules du sein absorbent davantage de calcium. Cette activité cellulaire accrue peut mettre sur la piste d'un cancer en train de se développer dans la même région [37] (voir la figure I.9).

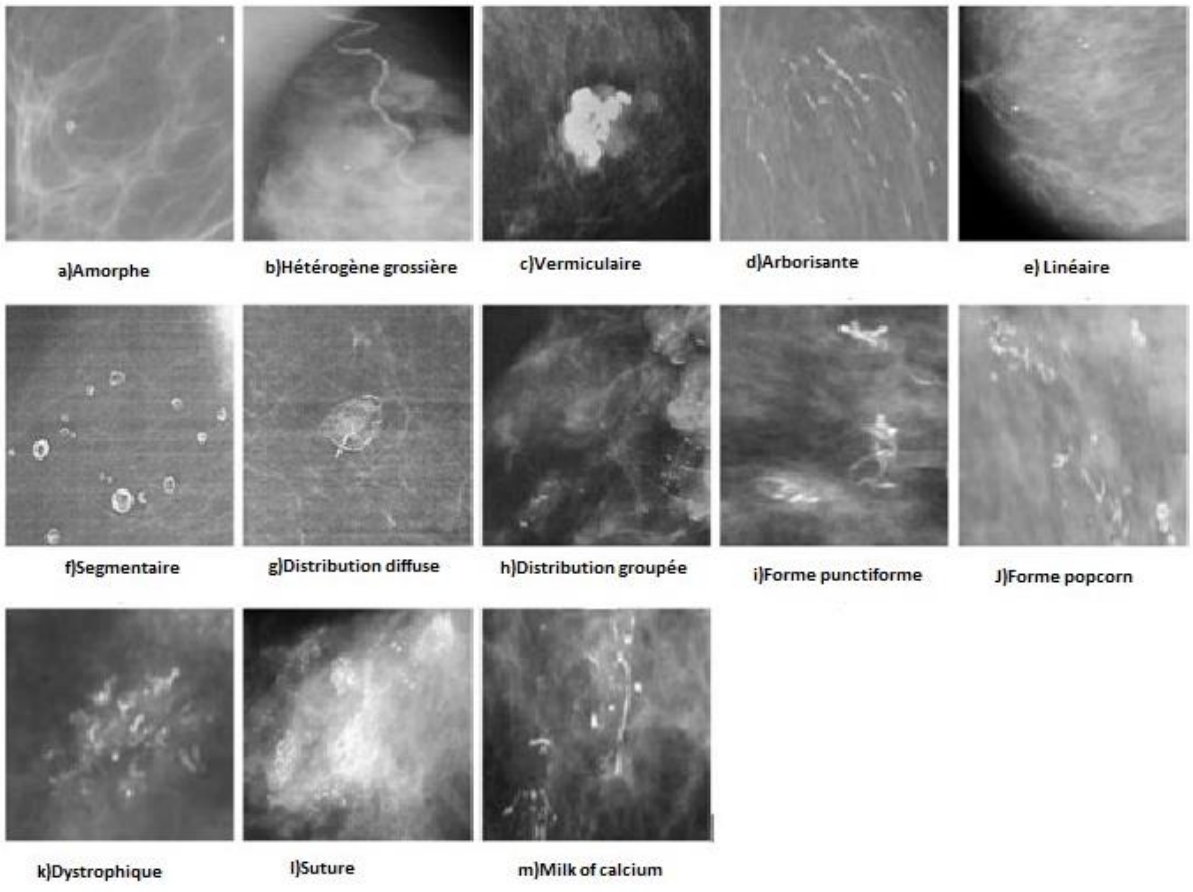


Figure 1.9 : différents types des micro-calcifications [37].

I.6.2. Les masses

Une masse est une opacité importante occupant un espace dans le sein et vue comme une tâche blanche sur l'image mammographie. Il peut s'agir d'un kyste (collection liquidienne non cancéreuse) ou d'une lésion solide, qui peut correspondre à un cancer de sein. Différents attributs permettent aux médecins de la décrire en vue de déterminer leur nature : sa forme, son contour et sa densité.

1. La forme

Selon la description du BIRADS, les masses mammaires peuvent avoir les formes suivantes [38] (voir la figure I.10) :

a. **Ronde** : Il s'agit de masse sphérique, circulaire ou globuleuse.
b. **Ovale** : Elle présente une forme elliptique (ou en forme d'œuf).
c. **Lobulée** : La forme de la masse présente une légère ondulation.
d. **Irrégulière** : Cette appellation est réservée aux masses dont la forme est aléatoire.

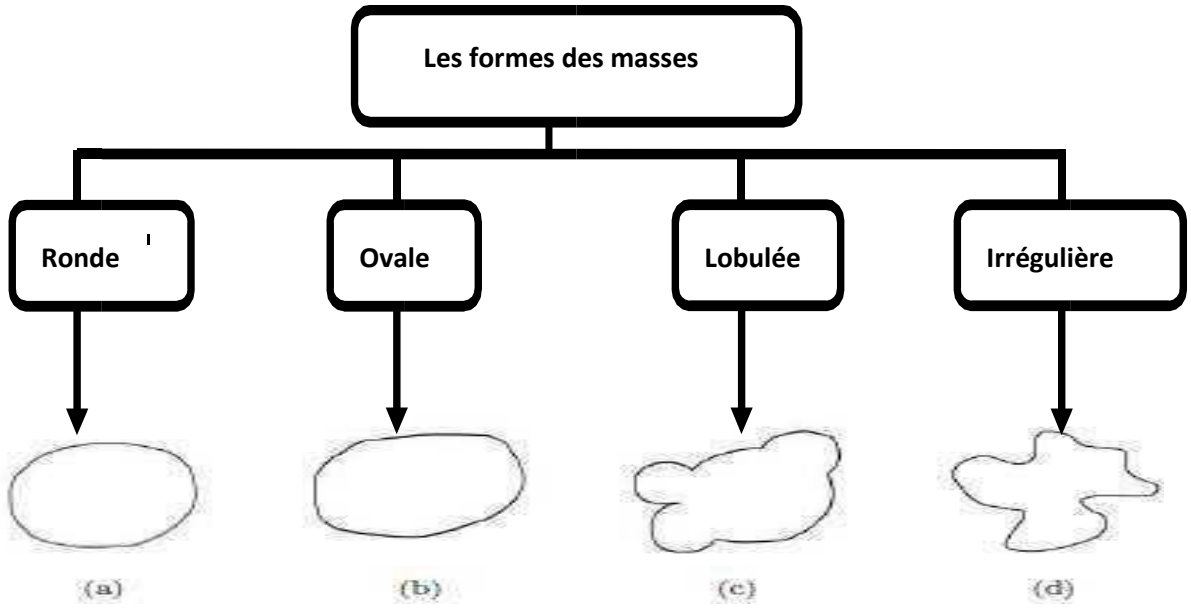


Figure 1.10 : Les différentes formes possibles d'une masse [39].

2. La densité

L'aspect du sein normal est très variable d'une femme à l'autre (voir la figure I.11). Le facteur le plus remarquable est la grande variabilité de la densité radiologique de l'aire mammaire. Les tissus mammaires (graisseux, conjonctifs et glandulaires) changent au fur et à mesure en vieillissant. Selon la densité du tissu mammaire est proportionnelle au risque de développement d'un cancer [40].

Afin de standardiser les comptes rendus mammographiques, la classification BIRADS de l'ACR à définit 4 classes de la composition du sein [41].

Les quatre types de la densité mammaire sont :

- Densité de type 1 : les seins sont sombres, presque entièrement graisseux avec un aspect homogène. Ils comptent moins de 25 % de tissu fibro-glandulaire, ce qui concerne 40 % des femmes.
- Densité de type 2 : de 25 à 50 % de tissu fibro-glandulaire, ce qui concerne 25 % des femmes.
- Densité de type 3 : de 50 à 75 % de tissu fibro-glandulaire, ce qui concerne 25 % des femmes.
- Densité de type 4 : les seins présentent un aspect extrêmement dense, homogène. Ils contiennent plus de 75 % de tissu fibro-glandulaire et concernent 10 % des femmes.

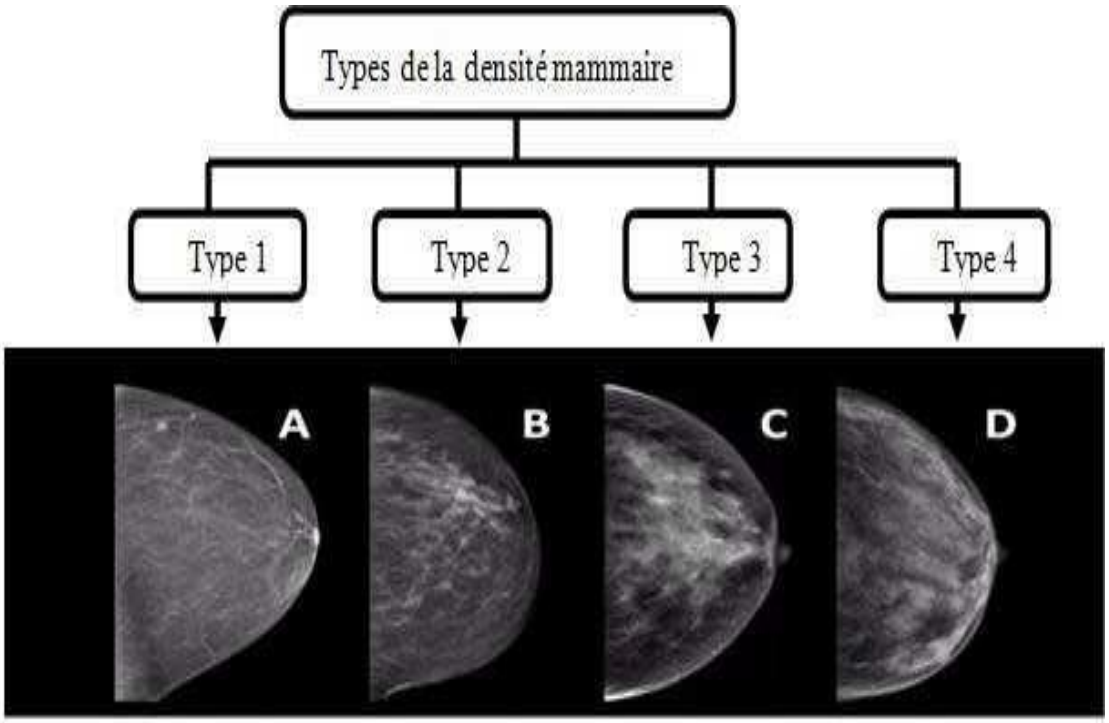


Figure I.11: Les différents types de densité mammaire [42].

3. Le contour

Le BIRADS dénombre cinq classes de contour des masses mammaires [43] (voir la figure I.12) :

a. Circonscrit : il s'agit d'une transition brusque entre la lésion et le tissu environnant.
b. Indistinct : dans ce cas, le contour est mal défini. Ce caractère indistinct (le contraire de circonscrit) peut correspondre à une infiltration.
c. Masqué : est un contour qui est caché par le tissu normal adjacent. Ce terme est employé pour caractériser une masse circonscrite dont une partie du contour est cachée.
d. Spéculé : la masse est caractérisée par des lignes radiaires prenant naissance sur le contour de la masse. Ces lignes radiaires sont appelées les spicules.
e. Micro-lobulé : dans ce cas, de courtes dentelures du contour créent de petites ondulations.
f. Distorsions architecturales : ce sont des signes hautement suspects qui se traduisent par des structures linéaires ou des spécules qui convergent vers une même zone focale.

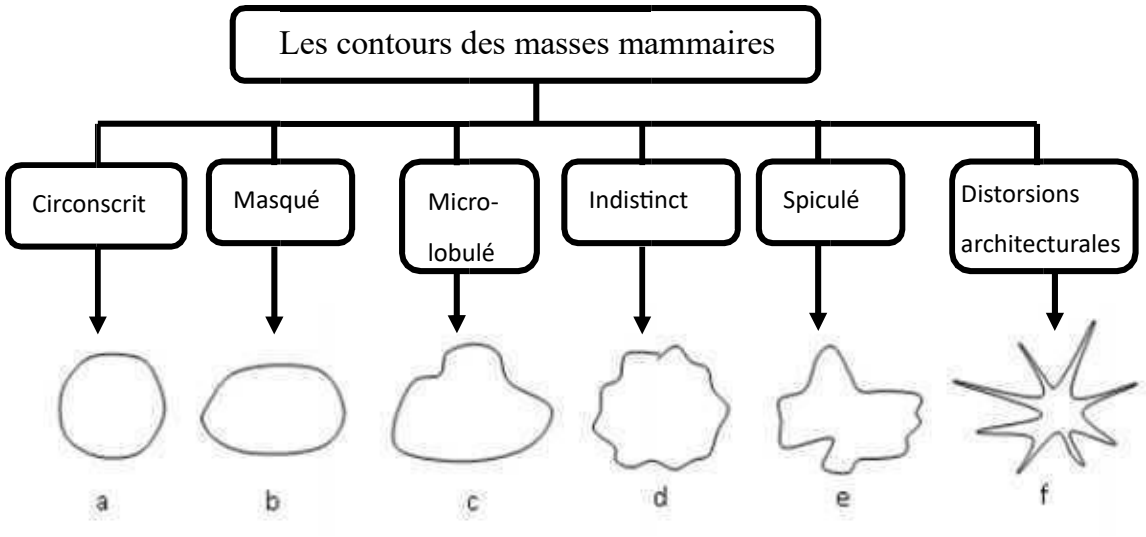


Figure I.12: Les différents contours d'une masse mammaire [44].

I.7. Diagnostic du cancer du sein assisté par ordinateur (CADx)

Le cancer du sein représente le type de cancer le plus courant chez les femmes, avec plus de deux millions de cas diagnostiqués chaque année. Afin de contribuer à accélérer le diagnostic, le projet SmartMamma CAD a mis au point de nouvelles méthodes de traitement d'images reposant sur le diagnostic assisté par ordinateur (CAD).

Le diagnostic des lésions mammaires par les radiologues est une tâche subjective et des erreurs peuvent donc être commises, dont les conséquences peuvent être fatales pour la patiente. La combinaison de la fatigue, du manque d'expérience en imagerie mammaire et du nombre d'images à analyser chaque jour rendait cette tâche fastidieuse et justifiait les erreurs de diagnostic. En revanche, les systèmes automatisés de diagnostic assisté par ordinateur (CADx) conçus sur des modèles prenant en compte les descripteurs liés aux lésions mammaires ont démontré leur robustesse, ce qui contribuera à réduire le nombre de biopsies inutiles. En effet, il a été démontré que moins d'un tiers des mammographies suspectes après biopsie sont en réalité des mammographies de lésions cancéreuses. De plus, les systèmes CADx sont conçus pour aider les radiologues à réduire le nombre de biopsies et à éviter le stress inutile du patient [45].

Ainsi, un système automatisé CADx utilisé dans un contexte de dépistage ou de diagnostic agirait comme un deuxième avis pour les radiologues et n'a absolument pas vocation à remplacer ces derniers. Typiquement, le fonctionnement d'un système CADx est une série d'étapes d'égale importance. Lorsque le système reçoit une image en entrée, il la prétraite

d'abord, c'est-à-dire qu'il supprime le bruit d'acquisition et améliore le contraste de l'image pour obtenir la meilleure qualité d'image possible [45].

Cette étape est suivie de la détection d'anomalies (telles que des micro calcifications, des opacités ou des distorsions structurelles). Plusieurs attributs mathématiques sont ensuite extraits des anomalies détectées pour mieux décrire la nature de la lésion. Enfin, les descripteurs extraits sont classés selon un algorithme adapté à la tâche pour déterminer la nature de la qualité [45].

I.7.1. L'importance des systèmes de diagnostic par ordinateur :

Même si une lésion est très suspecte à la mammographie ou à l'échographie, le diagnostic préopératoire est désormais indispensable pour deux raisons [45] :

- De nombreuses chirurgies inutiles pourraient être évitées s'il n'y avait pas le cancer. En cas de cancer du sein, la chirurgie peut être optimisée d'emblée : si le cancer est de petite taille, par exemple, intervention au niveau des ganglions lymphatiques (par ex ganglion sentinelle) peuvent être programmés et permettent au patient d'éviter le curage ganglionnaire axillaire. Par exemple, dans d'autres cas, le traitement peut commencer par des médicaments (thérapie néo adjuvante) pour permettre une chirurgie moins sévère plus tard.
- Dans de rares cas, ce diagnostic préopératoire n'est pas possible car la localisation radiographique, comme la biopsie, n'est pas disponible. La chirurgie permettra d'établir le diagnostic.

I.8. Les bases d'images de mammographie :

Il existe différentes bases d'images utilisées par les chercheurs en imagerie du sein pour concevoir des systèmes automatiques de diagnostic assisté par ordinateur (CADx). Parmi ces bases, on cite :

I.8.1. Mammographie Image Analysais Society (MIAS)

Mammographic Image Analysis Society (MIAS) est la plus ancienne base d'images publiquement disponible ; elle a été conçue au Royaume-Uni en 1994 et est encore largement utilisée dans l'état-de-l'art. La MIAS contient 161 cas pour un total de 322 images numérisées obtenues en incidence MLO, sur lesquelles sont présentes toutes les pathologies à savoir les lésions bénignes et malignes, mais également les images normales. Cette base dispose d'un nombre important de masses spéculées, ainsi que d'une information relative à la densité du sein ; toutefois, la classification des masses réalisée par les radiologues ayant évalué cette base ne

respecte pas les standards de l'ACR. Cependant, avec l'augmentation de l'usage de la classification BI-RADS de l'ACR, certains auteurs se sont essayés `a une classification des masses de la MIAS afin qu'elle puisse correspondre aux standards en vigueur en imagerie du sein.

Un autre défaut de la MIAS est l'annotation de ces images, qui consiste à indiquer le centre et le rayon de la région d'intérêt, c'est-`a-dire la région où se situe la pathologie. Ce genre d'annotation est considérée comme insuffisant pour certaines études comme la segmentation des masses, ou l'on souhaite que toutes les lésions circonscrites ou spéculées soient manuellement segmentées afin de faire une comparaison très précise avec les méthodes automatiques. Pour finir, l'autre inconvénient de la MIAS, c'est la résolution à laquelle les images sont numérisées et qui fait que cette base ne convient pas pour des expériences liées à la détection des micro-calcifications [46].

I.8.2. Digital Data base For Screening Mammography (DDSM)

Ce CBIS-DDSM (Curated Breast Imaging Subset of DDSM) est une version mise à jour et standardisée de la Digital Database for Screening Mammography (DDSM). La DDSM est une base de données de 2 620 mammographies sur films numérisés. Elle contient des cas normaux, bénins et malins, accompagnés d'informations anatomopathologiques vérifiées. L'ampleur de la base de données, ainsi que la validation des données de terrain, font de la DDSM un outil précieux pour le développement et le test de systèmes d'aide à la décision. La collection CBIS-DDSM comprend un sous-ensemble des données DDSM, sélectionné et organisé par un mammographe qualifié. Les images ont été décompressées et converties au format DICOM. La segmentation des régions d'intérêt (ROI) et les cadres de délimitation mis à jour, ainsi que le diagnostic pathologique pour les données d'apprentissage, sont également inclus [47].

Chaque étude comprend deux images de chaque sein, ainsi que des informations sur la patiente (âge au moment de l'étude, densité mammaire ACR, degré de subtilité des anomalies, description des anomalies par mots-clés ACR) et des informations sur l'image (scanner, résolution spatiale, etc.). Les images contenant des zones suspectes sont associées à des informations de référence au niveau du pixel concernant leur localisation et leur type. Un logiciel est également fourni pour accéder aux images mammographiques et de référence, ainsi que pour calculer les performances des algorithmes d'analyse d'images automatisés.

Les résultats de recherche publiés issus du développement de systèmes d'aide à la décision en mammographie sont difficiles à reproduire en raison de l'absence d'un ensemble de données

d'évaluation standardisé ; la plupart des algorithmes de diagnostic assisté par ordinateur (CADx) et de détection (CADe) du cancer du sein en mammographie sont évalués sur des ensembles de données privés ou sur des sous-ensembles non spécifiés de bases de données publiques. Peu d'ensembles de données publics bien organisés ont été mis à la disposition de la communauté mammographique. Parmi ceux-ci figurent la DDSM, la base de données de la Mammographic Imaging Analysis Society (MIAS) et le projet Image Retrieval in Medical Applications (IRMA). Bien que ces ensembles de données publics soient utiles, leur taille et leur accessibilité sont limitées [48].

I.8.3. Banco Web LAPIMO

La base de données BancoWeb LAPIMO est le nom d'un ensemble de femmes accessible en ligne construit par le (LAPIMO) de l'université de São Paulo (Brésil) pour les études dans le cadre d'applications du traitement d'images médicales, plus spécifiquement celle des mammographies. Elle contient plus de 1 400 images mammographiques liées à 320 cas cliniques, contenant plusieurs prises de vue (CC, MLO, agrandissements), annotées d'informations cliniques comme les catégories BI-RADS, la qualité de bénin é ou malignité des diagnostics, et la localisation spatiale des lésions. Les images d'origine sont fournies au format TIFF (8 - 16 bits), tandis que les images sont disponibles sous forme de pré manches JPEG pour la visualisation rapide. La banque de donné propose également des outils relatifs à la recherche, au découpage, au téléchargement et à la gestion des données. Son accès est conditionné par le fait d'être inscrit gratuitement sur le site officiel du projet suivi d'une demande d'autorisation soumise par courriel à l'administrateur (général, aux chercheurs ou professionnels de santé). BancoWeb est largement utilisé dans le cadre de projets pour la classification, la détection du cancer du sein, l'insertion d'intelligence artificielle, les systèmes d'aide au diagnostic, sur la base de la qualité de ses images et de la richesse de leurs annotations [49].

I.9. Conclusion

Dans ce chapitre, nous avons présenté les notions fondamentales relatives au cancer du sein et à l'imagerie mammaire, en particulier la mammographie. Après une brève description de l'anatomie de la glande mammaire, nous avons souligné l'importance de cette connaissance pour comprendre l'apparition et la localisation des anomalies mammaires.

L'objectif principal de cette introduction médicale était de mettre en évidence le rôle central de la mammographie dans la détection précoce du cancer du sein. À travers cette démarche,

nous insistons également sur la nécessité de sensibiliser à l'importance du dépistage régulier, car, dans le cas du cancer du sein, prévenir reste toujours préférable à guérir.

CHAPITRE : II

GÉNÉRALITÉS SUR MACHINE LEARNING

II.1 Introduction

L'évolution rapide des technologies numériques et de la puissance de calcul a favorisé l'émergence de domaines avancés de l'intelligence artificielle (IA), au sein desquels le Machine Learning (ML) occupe une place centrale. Le ML permet aux machines d'apprendre automatiquement à partir des données et de prendre des décisions sans qu'elles aient été explicitement programmées pour chaque tâche.

Aujourd'hui, apprentissage automatique (ML) est largement utilisé dans des secteurs variés tels que la santé, la finance, la reconnaissance faciale, ou encore la conduite autonome. Son potentiel en fait un outil puissant pour traiter des volumes massifs de données, extraire des modèles pertinents et automatiser des processus complexes.

Ce chapitre a pour objectif de présenter les fondements du ML, en proposant un aperçu historique, une classification des différentes approches, un panorama des méthodes les plus courantes ainsi que quelques exemples d'applications concrètes.

II.2 Historique

Depuis l'antiquité, le sujet des machines pensantes préoccupe les esprits. Ce concept est la base de pensées pour ce qui deviendra ensuite l'intelligence artificielle, ainsi qu'une de ses sous-branches : l'apprentissage automatique. La concrétisation de cette idée est principalement due à Alan Turing (mathématicien et cryptologue britannique) et à son concept de la « machine universelle » en 1936, qui est à la base des ordinateurs d'aujourd'hui. Il continuera à poser les bases de l'apprentissage automatique, avec son article sur « L'ordinateur et l'intelligence » en 1950, dans lequel il développe, entre autres, le test de Turing.

En 1943, le neurophysiologiste Warren McCulloch et le mathématicien Walter Pitts publient un article décrivant le fonctionnement de neurones en les représentant à l'aide de circuits électriques. Cette représentation sera la base théorique des réseaux neuronaux.

Arthur Samuel, informaticien américain pionnier dans le secteur de l'intelligence artificielle, est le premier à faire usage de l'expression machine learning (en français, « apprentissage automatique ») en 1959 à la suite de la création de son programme pour IBM en 1952. Le programme jouait au Jeu de Dames et s'améliorait en jouant. À terme, il parvint à battre le 4^{e} meilleur joueur des États-Unis.

Une avancée majeure dans le secteur de l'intelligence machine est le succès de l'ordinateur développé par IBM, Deep Blue, qui est le premier à vaincre le champion mondial d'échecs Garry Kasparov en 1997. Le projet Deep Blue en inspirera nombre d'autres dans le cadre de l'intelligence artificielle, particulièrement un autre grand défi : IBM Watson, l'ordinateur dont le but est de gagner au jeu Jeopardy. Ce but est atteint en 2011, quand Watson gagne à Jeopardy ! en répondant aux questions par traitement de langage naturel. Durant les années suivantes, les applications de l'apprentissage automatique médiatisées se succèdent bien plus rapidement qu'auparavant.

- En 2012, un réseau neuronal développé par Google parvient à reconnaître des visages humains.
- En 2014, 64 ans après la prédiction d'Alan Turing, le dialogueur Eugene Goostman est le premier à réussir le test de Turing en parvenant à convaincre 33 % des juges humains au bout de cinq minutes de conversation qu'il est non pas un ordinateur, mais un garçon ukrainien de 13 ans.
- En 2015, une nouvelle étape importante est atteinte lorsque l'ordinateur « AlphaGo » de Google gagne contre un des meilleurs joueurs au jeu de Go, jeu de plateau considéré comme le plus dur du monde.

En 2016, un système d'intelligence artificielle à base d'apprentissage automatique nommé Lip Net parvient à lire sur les lèvres avec un grand taux de succès [50].

II.3. Définition

Le ML est un sous-ensemble de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés pour accomplir une tâche spécifique. En d'autres termes, au lieu de suivre des instructions strictes préalablement définies, les systèmes de ML utilisent des algorithmes pour identifier des modèles et des relations dans des ensembles de données et font ensuite des prédictions ou prennent des décisions basées sur ces données. Ces machines améliorent, notamment la montée en compétences (voir la Figure II.1).

Contrairement aux systèmes traditionnels qui effectuent des tâches selon un programme codé à la main, le ML permet aux machines de s'améliorer progressivement en fonction de l'expérience, ce qui les rend plus adaptables et autonomes [51].

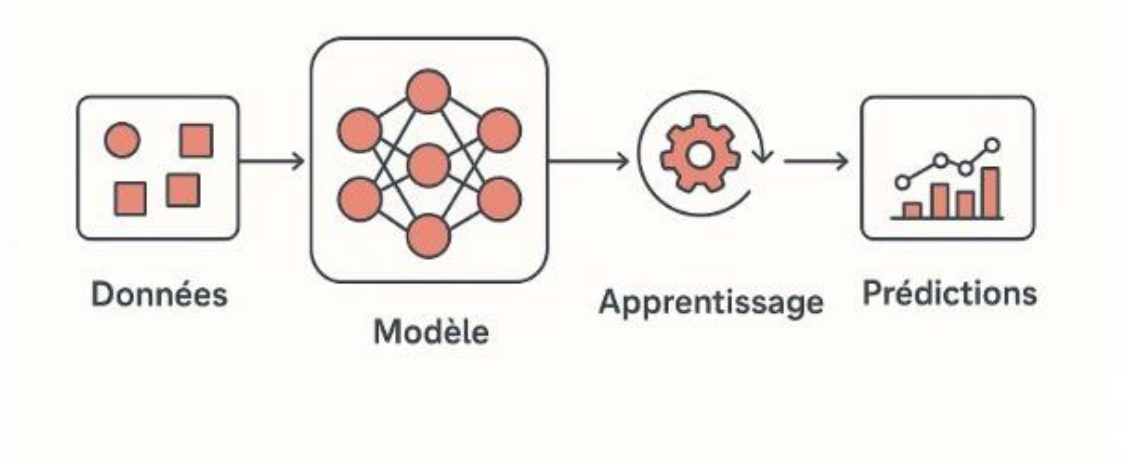


Figure II.1: Architecture générale de l'apprentissage automatique [52].

II.4. Les objectifs du machine learning

Les différents objectifs notables sont [53] :

- **La prédiction :** Créer des patterns (modèles) capables d'établir des prédictions précises sur de nouvelles données, tel est l'objectif du ML. En utilisant des ensembles de données d'entraînement, les algorithmes de ML peuvent analyser les relations entre les variables et construire des modèles prédictifs. Ces derniers peuvent être utilisés pour prédire des valeurs manquantes, estimer des résultats futurs ou anticiper des tendances.
- **La classification :** Les algorithmes de ML permettent également de classer les données dans des catégories prédéfinies. Comme vu précédemment, il peut être utilisé afin de placer des e-mails en spam ou en courriers légitimes. On l'utilise aussi pour identifier des fraudes dans les transactions financières ou pour diagnostiquer des maladies à partir de symptômes.
- **La reconnaissance de formes :** Le ML permet l'identification de motifs et de structures complexes dans les données. Nous pouvons citer la reconnaissance d'images, de la parole, la détection d'objets, la transcription automatique, la compréhension du langage naturel, etc. Ces capacités permettent de développer des systèmes intelligents capables de comprendre et d'interagir avec des données non structurées.
- **La recommandation :** La recommandation constitue assurément un grand objectif de travail auquel nous faisons souvent face. Les recommandations personnalisées envahissent de toute part les écrans qui sont les nôtres : ces algorithmes investissent les préférences et comportements d'un utilisateur, afin de lui proposer, à la façon de

l'Amazon de la grande distribution, des produits, des services ou des contenus correspondants à ses goût. D'une mise en œuvre importante, elle est devenue incontournable chez les e-commerçants, les plateformes de streaming ou les réseaux sociaux.

- **L'optimisation :** Le ML peut être utilisé pour améliorer des processus et des décisions. Par exemple, il peut aider à optimiser la gestion des stocks, l'affectation des ressources, la planification des itinéraires ou la tarification dynamique. En analysant les données historiques et en apprenant des modèles de performance, le Machine Learning peut aider à prendre des décisions plus efficaces et à améliorer les résultats.

Les objectifs cités ci-dessus démontrent la polyvalence de cette technologie. Son potentiel à résoudre des problèmes complexes est bluffant. Grâce à des algorithmes intelligents, il est capable d'automatiser les processus.

II.5. Principe de fonctionnement de l'apprentissage automatique

Le développement d'un modèle de ML repose sur quatre étapes principales. En règle générale, c'est un data scientiste qui gère et supervise ce procédé. La première étape consiste à sélectionner et à préparer un ensemble de données d'entraînement. Ces données seront utilisées pour nourrir le modèle de ML pour apprendre à résoudre le problème pour lequel il est conçu.

Les données peuvent être étiquetées, afin d'indiquer au modèle les caractéristiques qu'il devra identifier. Elles peuvent aussi être non étiquetées, et le modèle devra repérer et extraire les caractéristiques récurrentes de lui-même. Dans les deux cas, les données doivent être soigneusement préparées, organisées et nettoyées. Dans le cas contraire, l'entraînement du modèle d'apprentissage automatique peut induire un biais, ce qui affectera directement la fiabilité de ses prédictions futures.

La deuxième étape consiste à sélectionner un algorithme à exécuter sur l'ensemble de données d'entraînement. Le type d'algorithme à utiliser dépend du type et du volume de données d'entraînement et du type de problème à résoudre.

La troisième étape est l'entraînement de l'algorithme. Il s'agit d'un processus itératif. Des variables sont exécutées à travers l'algorithme, et les résultats sont comparés avec ceux qu'il aurait dû produire. Les « poids » et le biais peuvent ensuite être ajustés pour accroître la précision du résultat.

On exécute ensuite de nouveau les variables jusqu'à ce que l'algorithme produise le résultat correct. L'algorithme, ainsi entraîné, est le modèle de ML.

La quatrième et dernière étape est l'utilisation et l'amélioration du modèle. On utilise le modèle sur de nouvelles données, dont la provenance dépend du problème à résoudre. Par exemple, un modèle de ML conçu pour détecter les spams sera utilisé sur des emails [54] (voir la figure II.2).

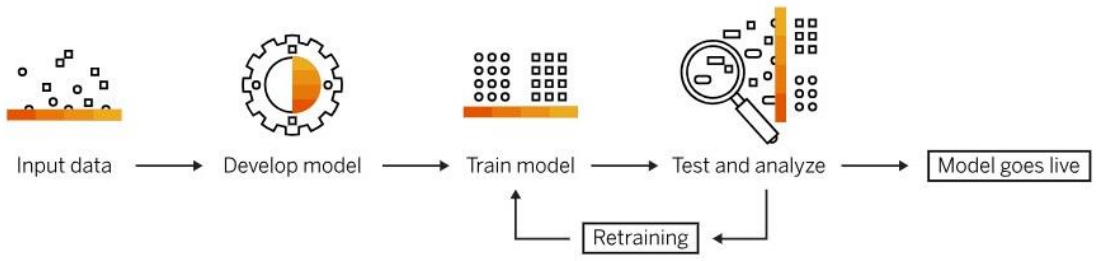


Figure II.2: Le fonctionnement du Machine Learning [55].

II.6. Les bases du machine learning

Pour bien comprendre le ML, il est nécessaire de connaître ses bases fondamentales. Voici les principaux éléments qui composent ce domaine [56] :

1. **Les données** : Le ML repose sur des données massives (big data) qui servent à entraîner les modèles. Ces données peuvent provenir de diverses sources telles que des images, du texte, des vidéos ou des capteurs. Plus les données sont variées et complètes, plus le modèle d'apprentissage sera performant.

2. **Les algorithmes** : Les algorithmes sont des suites d'instructions qui permettent aux machines de « comprendre » les données. Par exemple, un algorithme de régression linéaire peut être utilisé pour prédire une valeur continue, comme la température, en fonction de variables d'entrée.

3. **Les modèles** : Après avoir appris à partir des données, l'algorithme crée un modèle qui peut ensuite être utilisé pour faire des prédictions ou classifier des données nouvelles. Par exemple, un modèle formé pour détecter des emails de spam pourra identifier de nouveaux messages comme étant légitimes ou indésirables.

4. **L'évaluation et l'amélioration** : Une fois le modèle formé, il est évalué sur sa capacité à généraliser. Cela signifie qu'il doit être capable de faire des prédictions sur de nouvelles données qu'il n'a jamais vues auparavant. L'évaluation permet de déterminer si le modèle est efficace ou s'il nécessite des ajustements.

II.7. Les types d'apprentissage

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient :

II.7.1. Apprentissage supervisé

L'apprentissage supervisé est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés, au contraire de l'apprentissage non supervisé. On distingue les problèmes de régression des problèmes de classement. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification [57].

Les exemples annotés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée « hypothèse » ou « modèle ». On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et le but des méthodes d'apprentissage supervisé est de bien généraliser, c'est-à-dire d'apprendre une fonction qui fasse des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage (voir la figure II.3).

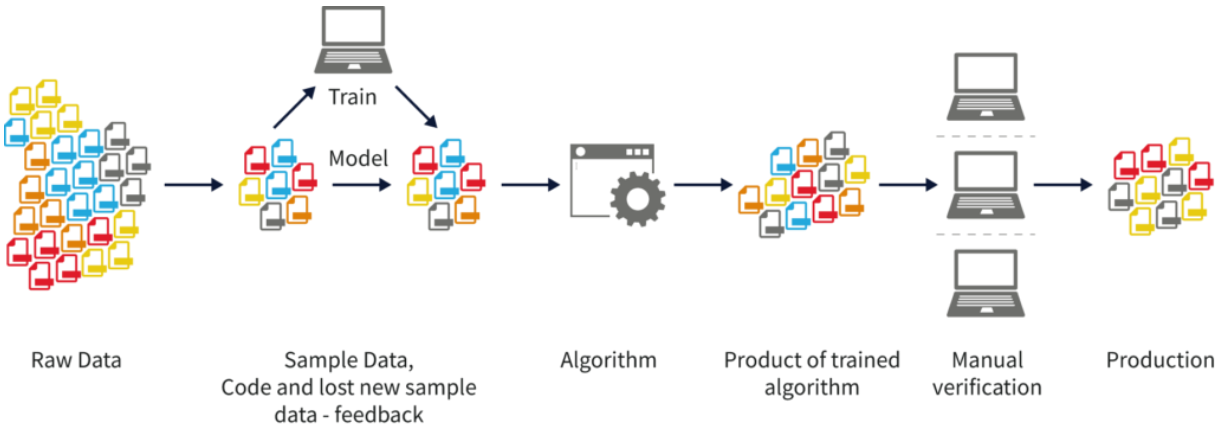


Figure II.3: Apprentissage supervisé [58].

II.7.2. Apprentissage non supervisé

Dans le domaine informatique et de l'intelligence artificielle, l'apprentissage non supervisé désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées (par exemple étiquetées comme « balle » ou « poisson »). Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées. Puisque les données ne sont pas étiquetées, il est impossible à l'algorithme de calculer de façon certaine un score de réussite. Ainsi, les méthodes non supervisées présentent une auto-organisation qui capture les modèles comme des densités de probabilité ou, dans le cas des réseaux de neurones, comme combinaison de préférences de caractéristiques neuronales encodées dans les poids et les activations de la machine [59].

En général, des systèmes d'apprentissage non supervisé permettent d'exécuter des tâches plus complexes que les systèmes d'apprentissage supervisé, mais ils peuvent aussi être plus imprévisibles. Même si un système d'IA d'apprentissage non supervisé parvient tout seul, par exemple, à faire le tri entre des chats et des chiens, il peut aussi ajouter des catégories inattendues et non désirées, et classer des races inhabituelles, introduisant plus de bruit que d'ordre (voir la figure II.4).

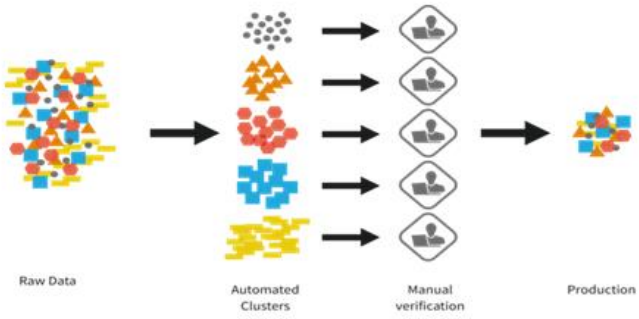


Figure II.4 : Apprentissage non supervisé [60].

II.7.3. Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage [61].

Un autre intérêt provient du fait que l'étiquetage de données nécessite souvent l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident.

Un exemple d'apprentissage semi-supervisé est le Co-apprentissag, dans lequel deux classifieurs apprennent un ensemble de données, mais en utilisant chacun un ensemble de caractéristiques différentes, idéalement indépendantes. Si les données sont des individus à classer en hommes et femmes, l'un pourra utiliser la taille et l'autre la pilosité par exemple [62].

II.7.4. Apprentissage auto-supervisé

L'apprentissage auto-supervisé (*self-supervised learning*, SSL) est une méthode d'apprentissage automatique où le modèle apprend à partir d'échantillons de données non annotées. Il peut être considéré comme une forme intermédiaire entre l'apprentissage supervisé et non supervisé. L'apprentissage auto-supervisé est typiquement utilisé sur des architectures à base de réseau de neurones artificiels. Le réseau de neurones apprend en deux étapes. Tout d'abord, la tâche est résolue sur la base de pseudo-étiquettes qui aident à initialiser les poids du réseau. Deuxièmement, la tâche réelle est effectuée avec un apprentissage supervisé ou non supervisé. L'apprentissage auto-supervisé a produit des résultats prometteurs ces dernières années et a trouvé une application pratique dans le traitement audio et est utilisé par Facebook et d'autres pour la reconnaissance vocale. Le principal attrait du SSL est que la formation peut se produire avec des données de qualité inférieure, plutôt que d'améliorer les résultats finaux. L'apprentissage auto-supervisé imite de plus près la façon dont les humains apprennent à classifier les objets [63] .

II.7.5. Apprentissage par renforcement

En intelligence artificielle, plus précisément en apprentissage automatique, l'apprentissage par renforcement consiste, pour un agent autonome (ex. : robot, agent conversationnel, personnage dans un jeu vidéo, etc.), à apprendre les actions à prendre, à partir d'expériences, de façon à optimiser une récompense quantitative au cours du temps. L'agent est plongé au sein d'un environnement et prend ses décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative. L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (appelé stratégie ou politique, et qui est une fonction associant à l'état courant l'action à exécuter) optimal, en ce sens qu'il maximise la somme des récompenses au cours du temps [64].

L'apprentissage par renforcement est l'une des trois grandes techniques d'apprentissage automatique, au côté de l'apprentissage supervisé et de l'apprentissage non supervisé [65].

II.7.6. Apprentissage par transfert

Le transfert d'apprentissage est une méthode du ML qui consiste à reprendre un modèle pré-entraîné sur une tâche A (source) pour l'adapter à une nouvelle tâche B (cible) apparentée afin d'améliorer ses performances sur cette tâche. Au lieu d'entraîner un modèle de zéro, ce qui coûte cher en donnée, calcul et temps, on fait appel aux connaissances acquises (features, représentations, régularités statistiques) d'une tâche antérieure pour faciliter l'apprentissage d'une nouvelle tâche, souvent avec un jeu de données plus restreint [66].

Précisément, on se base sur un modèle ayant reçu un pré-entrainement sur un énormément large corpus de données (ex. images, texte, audio) pour réaliser un peaufinage (fine-tuning) et l'adapter au cas particulier que l'on veut étudier en fonction de ses paramètres sur un jeu de données propre à ce nouveau problème à traiter. Cette méthode s'avère particulièrement performante en apprentissage profond, où les modèles doivent en général être alimentés par des ensembles de données conséquents pour fonctionner de manière adéquate [67].

II.8. Avantages et inconvénients du Machine Learning

❖ **Avantages**

L'apprentissage automatique accompagne de très nombreux bénéfices pour les entreprises et les différents secteurs car il est capable d'apprendre à partir de données sans programmation préalable. En voici les principales présentations [68] :

- ✓ **Automatisation de tâches répétitives :** Il permet d'automatiser des actions répétitives ce qui libère du temps et des ressources humaines pour des réalisations à plus forte valeur ajoutée et qui diminuent aussi les erreurs humaines.
- ✓ **Analyse de grandes quantités de données :** L'apprentissage automatique peut rapidement traiter d'énormes données à la recherche de modèles, de tendances et de corrélations cachées aux yeux des humains.
- ✓ **Prédictions de la précision et en amélioration dans le temps** : Il s'améliore au temps de recherche des prévisions de toute matière dans de nombreux domaines comme la finance, la supply chain, la maintenance industrielle, ou le comportement des consommateurs par exemple.

- ✓ **Personnalisation des statistiques utilisateurs** : Il facilite les interactions au vu du comportement et préférences individuels des uns et des autres, et contribue à la maintenance de l'engagement, de la satisfaction et fidélité des consommateurs.
- ✓ **Optimisation de la production des opérations opérationnelles** : Il identifie les faiblesses, goulots d'étranglement et opportunités d'amélioration et contribue à l'optimisation des performances de systèmes, réseaux et opérations opérationnelles internes.
- ✓ **Repérage et Prévention des fraudes** : Son usage a pour effet d'accélérer la détection des fraudes, donc de renforcer la sécurité des entreprises via des réponses rapides.
- ✓ **Innovation et développement de nouveaux produits :** Il va permettre de créer des solutions innovantes comme les assistants virtuels intelligents ou les systèmes de recommandation avancés.

En résumé, le ML est un facteur clé dans l'amélioration des décisions basées sur les données pour faire croître, trouver de nouveaux leviers de revenus ou maximiser la performance business.

❖ **Inconvénients**

Les inconvénients majeurs du ML sont les suivants [68] :

- ✓ **Nécessité de grandes quantités de données de haute qualité** : Pour donner des résultats fiables, le machine learning a besoin d'ensembles de données volumineux, complets, non biaisés. Or, la collecte, le nettoyage et la préparation de ces données peuvent s'avérer coûteux, ou longs. Les biais sur les données présentent le risque de produire des résultats faussés, aggravant éventuellement les discriminations ou les erreurs déjà existantes.
- ✓ **Coûts élevés et besoins importants en calcul** : Les ressources matérielles nécessaires sont souvent conséquentes, notamment l'investissement initial (en matériel – GPU, serveurs – et infrastructure) pour l'entraînement de modèles complexes – surtout en deep learning, qui requiert une importante puissance de calcul. Le coût de cette technologie en limite souvent l'accès aux seules grandes entreprises, voire aux centres de recherche.
- ✓ **Complexité de mise en œuvre** : La mise en œuvre de la solution requiert des compétences techniques élevées. Il est souvent nécessaire de recruter des spécialistes en

data science et de former les équipes, ce qui constitue un frein, notamment pour les PME.

- ✓ **Difficulté d'interprétation des modèles** : Beaucoup de modèles, en particulier les réseaux de neurones profonds, sont des "boîtes noires". Ces modèles délivrent des résultats sans expliquer clairement leurs décisions, ce qui pose problème dans les secteurs où la transparence est indispensable, comme la finance ou la santé.
- ✓ **Risques liés à la confidentialité et à la sécurité des données** : L'utilisation de données sensibles fait naitre des enjeux cruciaux de protection de la vie privée. Il est impératif que les entreprises sécurisent ces données et respectent la réglementation [aussi bien au niveau du RGPD (Règlement Général sur la Protection des Données)].
- ✓ **Processus long et sujet aux erreurs** : L'entraînement requiert un temps considérable ; les algorithmes sont susceptibles de se tromper lorsque, d'une part, ils sont mal adaptés au problème traité et, d'autre part, alimentés par des données non conformes ou de mauvaise qualité. Les réglages et les méthodes d'optimisation des modèles sollicitent de multiples itérations.
- ✓ **Limites sur certaines tâches complexes** : Les résultats peuvent être restreints sur des tâches très complexes ou mal définies. Le ML ne peut pas s'appliquer de manière appropriée dans beaucoup d'applications pour les prédictions, à l'échelon mécanique, dans beaucoup de cas de figure.

En résumé, bien que très puissant, le ML nécessite des moyens techniques, humains et financiers importants, pose des défis éthiques et de transparence, et repose fortement sur la qualité des données pour être efficace.

II.9. Domaines d'application du Machine Learning

Les champs d'application du ML sont très larges, presque tous les secteurs d'activité sont touchés, c'est-à-dire [69] :

- **Secteur de la santé** : le ML est très utilisé pour réaliser des diagnostics sur la base d'images médicales, prédire le risque de survenance de pathologies par la mise en relation des courbes d'analyse du dossier patient, personnaliser le traitement en optimisation de la médecine prédictive (Anticipation du risque de rechute des patients ou des patientes atteints d'un cancer, par exemple).
- **La finance :** est un domaine qui recourt aux modèles de ML afin de repérer les éventuelles fraudes, gérer des portefeuilles d'investissement, anticiper et prévoir les

tendances des marchés financiers et concevoir des stratégies d'investissement. Ils permettent aussi d'automatiser la conformité, d'offrir des services bancaires personnalisés, ou de gérer les transactions.

- **E-commerce et marketing** : Le ML est également à la base des systèmes de recommandation personnalisés (les fameux algorithmes qui font que vous aimez, à la fois, Amazon et Netflix), de la compréhension des comportements d'achats, de l'optimisation des campagnes publicitaires, de la sentiment analysis des clients.
- **Industrie et automatisation** : Le ML intervient dans l'optimisation des processus de production, la maintenance prédictive (anticipation des pannes machines), le contrôle qualité des chaînes de production ou la robotique intelligente dans les usines.
- **Sécurité :** Enfin, en matière de sécurité, on note le recours aux algorithmes de ML en reconnaissance faciale, en détection d'intrusion sur le réseau, en analyse vidéo-surveillance ou encore pour la cybersécurité (mis en œuvre pour détecter des comportements anormaux).
- **Automobile** : Prédictions pour la compréhension de l'environnement (capteur visuel webcam, décision temporelle de conduite assistée, modèle d'intelligence artificielle).
- **Agriculture** : Prédictions des rendements, diagnostics de maladies et nuisibles, optimisation des ressources en eau et engrais, analyse des données météorologiques pour la récolte.
- **Energie :** Planification de la consommation d'électricité, optimisation de la production et de la distribution (smart grid) préventive, maintenance des réseaux, efficacité énergétique.
- **Service client** : Chatbot, assistants virtuels, analyse des sentiments en vue de répondre, automatisation du suivi des tickets de support.
- **Reconnaissance d'image, de texte et de voix** : Reconnaissance d'images (médicale, sécurité, filtrage automatique de photos), reconnaissance vocale (assistants vocaux, transcription), traitement automatique du langage naturel pour le traducteur automatique ou l'analyse de documents.
- **Secteur public et sciences de la vie :** Détection de fraudes, amélioration de la gestion des ressources, aide à la politique publique et analyse de grand ensemble de données scientifiques.

- **La prédiction :** Créer des patterns (modèles) capables d'établir des prédictions précises sur de nouvelles données, tel est l'objectif du ML. En utilisant des ensembles de données

d'entraînement, les algorithmes de ML peuvent analyser les relations entre les variables et construire des modèles prédictifs. Ces derniers peuvent être utilisés pour prédire des valeurs manquantes, estimer des résultats futurs ou anticiper des tendances.

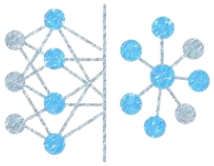
- **La classification :** Les algorithmes de ML permettent également de classer les données dans des catégories prédéfinies. Comme vu précédemment, il peut être utilisé afin de placer des e-mails en spam ou en courriers légitimes. On l'utilise aussi pour identifier des fraudes dans les transactions financières ou pour diagnostiquer des maladies à partir de symptômes.
- **La reconnaissance de formes :** Le ML permet l'identification de motifs et de structures complexes dans les données. Nous pouvons citer la reconnaissance d'images, de la parole, la détection d'objets, la transcription automatique, la compréhension du langage naturel, etc. Ces capacités permettent de développer des systèmes intelligents capables de comprendre et d'interagir avec des données non structurées.
- **La recommandation :** Il s'agit sans doute de l'objectif auquel nous sommes le plus exposés. Les recommandations personnalisées envahissent nos écrans. Ces algorithmes analysent les préférences et les comportements des utilisateurs pour leur suggérer des produits, des services ou des contenus pertinents. Ils sont couramment utilisés en e-commerce, par les services de streaming et les réseaux sociaux.
- **L'optimisation :** Le ML peut être utilisé pour améliorer des processus et des décisions. Par exemple, il peut aider à optimiser la gestion des stocks, l'affectation des ressources, la planification des itinéraires ou la tarification dynamique. En analysant les données historiques et en apprenant des modèles de performance, le Machine Learning peut aider à prendre des décisions plus efficaces et à améliorer les résultats.

Ces objectifs des différents modèles de ML démontrent la polyvalence de cette technologie. Son potentiel à résoudre des problèmes complexes est bluffant. Grâce à des algorithmes intelligents, il est capable d'automatiser les processus, de définir des insights à partir de données et de prendre des décisions basées sur des informations fiables.

II.10. Conclusion

Ce chapitre a présenté un panorama clair, structuré, allant des origines au présent du domaine du Machine Learning : le défi que représente la saisie des fondements théoriques de ce champ au travers

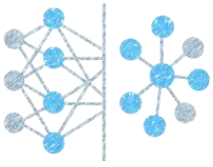
de ses types d'apprentissage (supervisé, non supervisé, semi-supervisé, auto-supervisé, par renforcement, par transfert), mais aussi des étapes clés de conception d'un modèle intelligent.



CHAPITRE : III

MÉTHODES DE CLASSIFICATION :

KNN ET ANN



III .1 Introduction

La classification est une technique très utilisée dans les travaux de fouille de données, la recherche d'information, reconnaissances des formes, le diagnostic médical, apprentissage automatique, l'aide à la décision et dans plusieurs domaines de recherche de l'intelligence artificielle.

Dans le cadre de ce mémoire, notre objectif est d'explorer et d'évaluer les performances de différents classifieurs issus de l'apprentissage supervisé, afin d'identifier ceux les plus adaptés à la classification d'anomalies mammaires à partir d'images médicales. C'est pourquoi, dans ce chapitre, nous avons décidé de commencer par la mise en œuvre et l'évaluation de deux classifieurs à savoir : le K-Nearest Neighbors (KNN) et Artificial Neural Network (ANN) en laissant la possibilité d'élargir l'étude à d'autres classifieurs dans des travaux futurs.

III.2. Méthodes de Classification

III.2.1. Concepts et Définitions :

La classification est le processus, qui permet de grouper des objets (observations ou individus) dans des classes (clusters) de manière à ce que les objets appartenant à la même classe soient plus similaires entre eux qu'aux objets appartenant aux autres classes. Le calcul de la proximité entre objets se fait sur une série de variables mesurées sur tous les objets. La classification connait une large utilisation dans plusieurs domaines notamment de l'intelligence artificielle comme l'analyse financière (prévision d'évolution de marchés), Marketing (établir un profil client, mailing), Banque (attribution de prêts), Médecine (aide au diagnostic), Télécom (détection de fraudes). Biométrie, Robotique, Reconnaissance de forme (OCR, Transcription de la parole, Compréhension/Dialogue), Recherche d'information (moteur internet, moteur multimédia) [70].

III.2.2. L'architecture typique d'une application basée sur la classification :

Comme la classification est un problème central en reconnaissance des formes, l'application de cette dernière lors des travaux de développement d'un outil de classification dans n'importe quel domaine, doit être anticipé par d'autres phases d'extraction et d'analyse des informations à classifier, ces travaux se sont également attachés à étudier les phases montrées dans la figure (voir la figure III.1) [70].

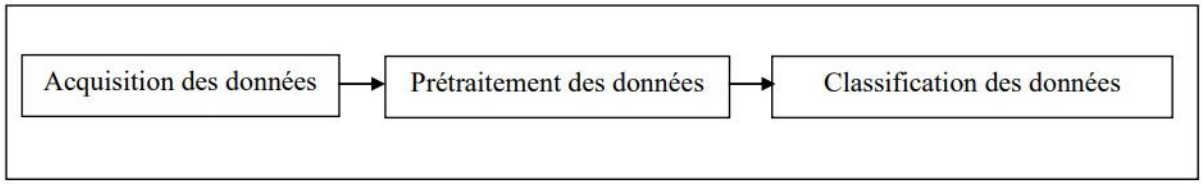


Figure III.1 : Étapes de la classification des données [70].

- **Acquisition des données** : d'une manière générale, il s'agit à ce niveau de mettre en place l'ensemble d'instrumentation de façons à reproduire le phénomène observé le plus fidèlement possible. C'est l'opération de transformation de l'information à traiter en signaux numériques manipulables par ordinateurs ou bien la numérisation de l'information.
- **Prétraitement des données** : Cette phase correspond au filtrage des informations en ne conservant que ce qui est pertinent dans le contexte d'étude.
- **Classification des données :** Elle correspond à l'étape de décision et pour cela plusieurs méthodes se présentent pour la résolution.

III.2.3. Taxonomie de la classification

Un grand nombre de méthodes est établi pour résoudre à peu près tous les problèmes de classification, cependant du fait que certaines de ces approches partagent des caractéristiques communes, soit dans la façon d'appréhender le problème (apprentissage supervisé ou non), soit dans la nature de la sortie réalisée (groupes disjoints ou classification flou). (Voir la figure III.2) montre le regroupement de ces méthodes sous la forme d'une hiérarchie (taxonomie) [70].

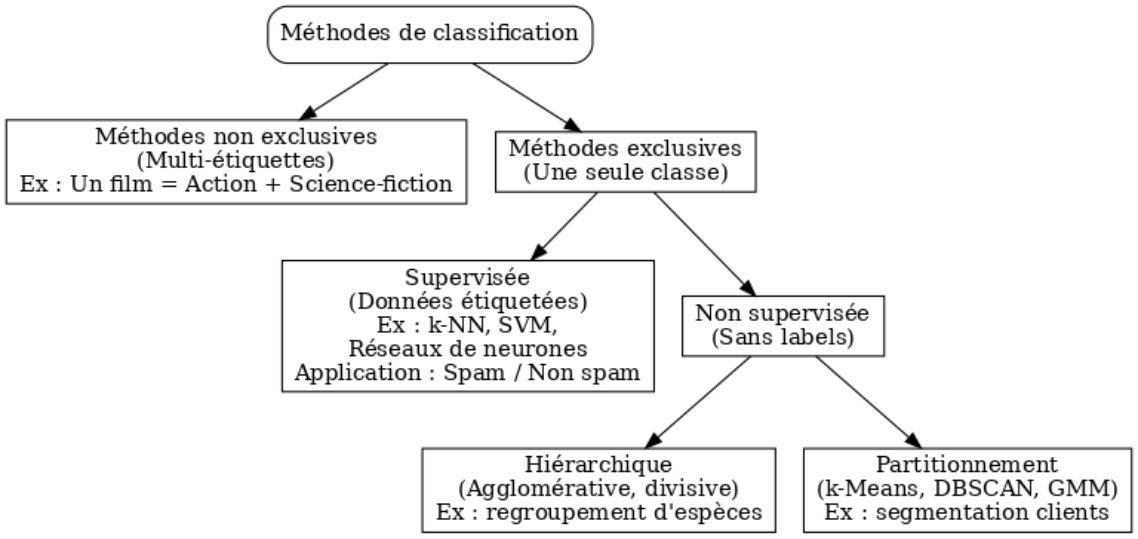


Figure III.2 : Les méthodes de classification [70].

III.3. Algorithme de KNN (K plus proches voisins)

III.3.1. Historique

L'algorithme des K plus proches voisins, le KNN (K-Nearest Neighbors), est sans doute l'un des plus anciens et des plus simples des algorithmes d'apprentissage supervisé. Son histoire débuta dès les années 1950 quand Evelyn Fix et Joseph Hodges, employés de l'US Air Force dans le cadre du projet militaire, proposèrent une méthode de classification non paramétrique, qu'ils appelèrent « discriminant analysis » et fondée sur le principe d'attribuer à un point inconnu la catégorie dans laquelle se rangent ses plus proches voisins dans l'espace des données. Ce travail, d'un accès confidentiel et qui ne fut alors pas publié, se voit précipité de répondre, en 1967, les travaux de Thomas Cover et Peter Hart illustrés par leur article « Nearest Neighbor Pattern Classification » apportait une preuve théorique expliquant que l'approche est efficace et fondant l'algorithme. Dans les années 1980, des améliorations sont mises en œuvre, avec la version floue notamment présentée par Adam Jozwik en 1983 et James Keller en 1985, qui viendront réduire le taux d'erreurs de la classification. De nos jours, le KNN est présenté comme un outil du fait de sa simplicité et de sa robustesse, utilisé dans de très nombreux champs comme la reconnaissance d'images, la bio-informatique ou la recommandation de contenus [71].

III.3.2. Concept et Définition

L'algorithme des K plus proches voisins, est un classificateur d'apprentissage supervisé non paramétrique, qui utilise la proximité pour effectuer des classifications ou des prédictions sur le regroupement d'un point de données individuel. Il est généralement utilisé comme algorithme de classification en partant de l'hypothèse que des points similaires peuvent être trouvés les uns à côtés des autres. Cependant, avant qu'une classification puisse être faite, la distance doit être définie. La distance euclidienne est la plus couramment utilisée, que nous aborderons plus en détail ci-dessous. Il convient également de noter que l'algorithme KNN fait également partie d'une famille de modèles "d'apprentissage paresseux", ce qui signifie qu'il ne stocke qu'un ensemble de données d'entraînement au lieu de subir une étape d'entraînement. Cela signifie également que tous les calculs ont lieu lorsqu'une classification ou une prédiction est effectuée. Puisqu'il s'appuie fortement sur la mémoire pour stocker toutes ses données d'entraînement, il est également appelé méthode d'apprentissage basée sur les instances ou basée sur la mémoire [71].

III.3.3. Le fonctionnement de l'algorithme de KNN

L'algorithme KNN est un algorithme d'apprentissage automatique qui peut être utilisé pour la classification, la régression et le regroupement. L'algorithme fonctionne en trouvant les k-voisins les plus proches d'un point de données donné, puis en classant ou en prédisant l'étiquette de ce point de données en fonction des étiquettes de ses voisins. La valeur de k est un hyper paramètre qui peut être ajusté pour optimiser les performances de l'algorithme. Il est conçu pour effectuer la tâche efficacement en sélectionnant la valeur optimale du k à partir des résultats obtenus [71].

III.3.4. Etapes de l'algorithme de KNN

1. Tout d'abord, l'algorithme prend en entrée un nouvel échantillon de données de test pour lequel la classe doit être prédite.
2. Ensuite, l'algorithme calcul la distance entre cet exemple et tous les exemples de la base de données d'entraînement, généralement en utilisant une mesure de distance telle que la distance euclidienne.
3. L'algorithme sélectionne les K exemples les plus proches du nouvel exemple en termes de distance. Ces exemples sont appelés les K plus proches voisins.
4. Pour la classification, l'algorithme attribue la classe majoritaire parmi les K plus proches voisins au nouvel exemple. Pour la régression, l'algorithme prédit la moyenne des valeurs des K plus proches voisins.
5. Enfin, l'algorithme renvoie la classe prédite ou la valeur prédite pour le nouvel exemple.

La valeur de K est un paramètre important de l'algorithme qui doit être spécifiée avant l'exécution de l'algorithme. Si K est trop petit, la décision sera sensible aux bruits dans les données, tandis que si K est trop grand, la décision sera trop générale et ne pourra pas détecter les subtilités des données. L'algorithme KNN est utilisé dans de nombreuses applications telles que la classification d'images, la reconnaissance de caractères manuscrits, la prédiction des prix de l'immobilier [72] (voir la figure III.3).

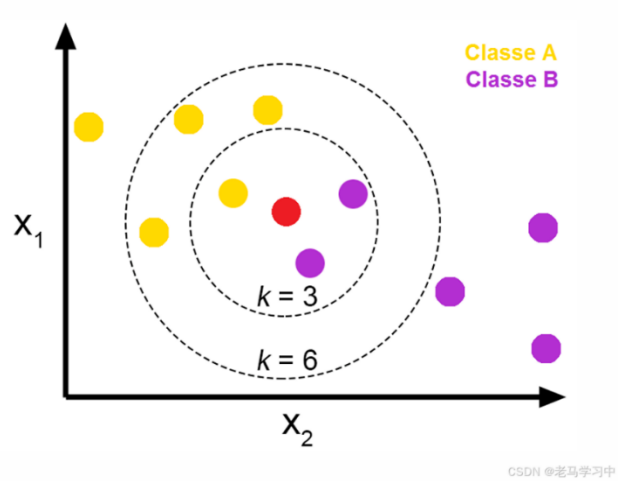


Figure III.3 : Fonctionnement de l'algorithme KNN [72].

Afin de déterminer quels points de données sont les plus proches d'un point de requête donné, la distance entre le point de requête et les autres points de données devra être calculée. Ces métriques de distance aident à former des limites de décision, qui partitionnent les points de requête en différentes régions.

III.3.5. Métriques de distance dans un algorithme KNN

Pour déterminer quels points de données sont les plus proches d'un point donné, il est nécessaire de calculer la distance entre ce point et les autres points de données. Ces mesures de distance aident à définir des frontières décisionnelles, qui partitionnent les points en différentes régions. Les frontières décisionnelles sont souvent représentées.

Bien qu'il existe plusieurs mesures de distance parmi lesquelles vous pouvez choisir, cet article ne couvre que les suivantes [73] :

- **Distance euclidienne** : il s'agit de la mesure de distance la plus couramment utilisée, et elle est limitée aux vecteurs à valeurs réelles. En utilisant la formule ci-dessous, il mesure une ligne droite entre le point de requête et l'autre point mesuré.

$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}} \qquad \textbf{Eq III.1}$$

x, y : vecteurs.

n : dimension des vecteurs x, y.

x_i : ième composante du vecteurs x.

y_i : ième composante du vecteur y.

- Exemple de distance euclidienne (voir la figure III.4)**:**

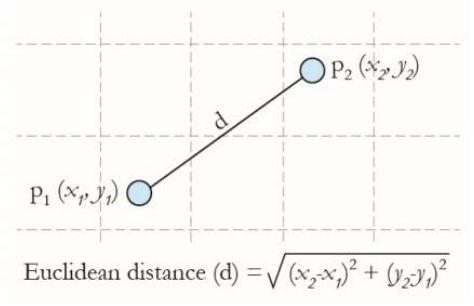


Figure III.4 : Exemple de Distance Euclidienne [73].

- **Distance de Manhattan :** il s'agit également d'une autre mesure de distance populaire, qui mesure la valeur absolue entre deux points. Elle est également appelée distance en taxi ou distance d'un pâté de maisons, car elle est généralement visualisée avec une grille, illustrant comment on peut naviguer d'une adresse à une autre via les rues de la ville.

$$\boldsymbol{d(x,y) = \sum_{i=1}^{n} |(x_i - y_i|} \qquad \textbf{Eq III.2}$$

xi et yi : les coordonnées des points x*x* et y*y* à la dimension i .

m : le nombre de dimensions.

- **Distance de Hamming** : Cette technique est généralement utilisée avec des vecteurs booléens ou de chaîne, identifiant les points où les vecteurs ne correspondent pas. En conséquence, il a également été appelé la métrique de chevauchement. Ceci peut être représenté par la formule suivante :

$$\boldsymbol{de(x,y) = \sum_{i=1}^{K} |xi - yi|} \qquad \textbf{Eq III.3}$$

$$\begin{cases} x = y & d_h = |xi - yi| = 0 \\ x \neq y & d_h \neq |xi - yi| \neq 1 \end{cases}$$

x et y : les deux vecteurs à comparer.

k: la longueur des vecteurs (nombre de positions).

xi et yi : les valeurs (0 ou 1) des vecteurs aux positions i.

d_h : la distance partielle à chaque position.

III.3.7. Applications de k-NN dans la machine learning

L'algorithme KNN est utilisé dans diverses applications, principalement dans le cadre de la classification. Voici quelques exemples [73] :

- **Prétraitement des données :** les jeux de données contiennent souvent des valeurs manquantes, mais l'algorithme KNN peut estimer ces valeurs via un processus appelé imputation des données manquantes.

- **Finance :** le KNN est également utilisé dans divers cas d'utilisation en finance et en économie. Par exemple, une publication montre comment KNN appliqué aux données de crédit peut aider les banques à évaluer les risques associés à un prêt. Il est également utilisé pour déterminer la solvabilité des demandeurs de prêt. Un autre article souligne son utilisation pour les prévisions boursières, les taux de change, les opérations à terme et les analyses de blanchiment d'argent.

- **Santé :** le KNN a des applications dans le domaine de la santé, notamment pour prédire les risques d'infarctus du myocarde ou de cancer de la prostate, en calculant les expressions de gènes les plus probables.

- **Reconnaissance de formes :** le KNN est également utilisé dans la reconnaissance de formes (ou de motifs), notamment pour la classification de chiffres et de textes. Il est particulièrement utile pour identifier des chiffres manuscrits sur des formulaires ou des enveloppes postales.

III.3.8. Avantages et Inconvénients de l'algorithme KNN

Comme tout algorithme de machine learning, KNN présente des avantages et des inconvénients. En fonction du projet et de l'application, il peut être ou non le bon choix [74].

❖ **Avantages**

- ✓ **Facile à mettre en œuvre :** en raison de sa simplicité et de sa précision, le KNN est l'un des premiers classificateurs appris par les débutants en science des données.
- ✓ **S'adapte facilement :** au fur et à mesure que de nouveaux échantillons d'entraînement sont ajoutés, l'algorithme s'ajuste pour inclure ces nouvelles données, car toutes les données d'entraînement sont conservées en mémoire.

- ✓ **Peu d'hyperparamètres :** KNN ne nécessite qu'une valeur pour k et une mesure de distance, ce qui le rend relativement simple par rapport à d'autres algorithmes de machine learning.

❖ **Inconvénients**

- ✓ **N'est pas très évolutif :** KNN étant un algorithme paresseux, il consomme davantage de mémoire et d'espace de stockage par rapport à d'autres classificateurs, ce qui peut s'avérer coûteux en temps et en argent. Plus de mémoire et de stockage entraînent des coûts supplémentaires pour l'entreprise, et le traitement de volumes de données plus importants peut prendre plus de temps. Bien que différentes structures de données, telles que Ball-Tree, aient été développées pour atténuer les inefficacités de calcul, un autre classificateur peut être plus adapté en fonction du problème à résoudre.
- ✓ **La malédiction de la dimensionnalité :** l'algorithme KNN a tendance à être victime de la malédiction de la dimensionnalité, ce qui signifie qu'il fonctionne mal avec des données d'entrée comportant de nombreuses dimensions. Ce phénomène est parfois appelé « phénomène de pic », où, après avoir atteint un nombre optimal de caractéristiques, l'ajout de nouvelles caractéristiques augmente les erreurs de classification, en particulier avec des échantillons de petite taille.
- ✓ **Tendance au sur-ajustement :** en raison de cette malédiction de la dimensionnalité, le KNN est également plus enclin au surajustement. Bien que des techniques de réduction de la dimensionnalité et de sélection des caractéristiques puissent être employées pour prévenir ce problème, la valeur de k a également un impact significatif sur le comportement du modèle. Des valeurs faibles de k peuvent entraîner un surajustement, tandis que des valeurs plus élevées ont tendance à « lisser » les prédictions en moyenne, mais si k est trop élevé, le modèle risque de sous-ajuster les données.

III.4. Réseaux de Neurones Artificiels (ANN)

Les réseaux de neurones artificiels (Artificial Neural Networks, ANN) permettent de reproduire, de manière formelle, certains mécanismes du cerveau humain. Les chercheurs ont ainsi pu mettre en évidence des similitudes avec le fonctionnement cérébral, notamment en termes de capacité d'apprentissage, de mémorisation, de reconnaissance des objets et de prise de décision. Comme nous le savons, le cerveau humain est constitué de milliards de neurones

interconnectés entre eux par des cellules neuronales, formant ainsi un immense réseau complexe d'unités associées.

Cette corrélation entre les cellules nerveuses donne à ces dernières la capacité de stocker et fournir des informations, des images, audio et succession de signaux qui reçoivent à travers les différents neurones, les réseaux de neurones permettent également d'apprendre par la répétition et l'erreur.

III.4.1. Concepts et Définitions

- **Historique**

L'histoire des réseaux de neurones artificiels revient au1943, où Mac Cullochet Pittsont proposé des neurones formels mimant les neurones biologiques et capables de mémoriser des fonctions booléennes simples. Les réseaux de neurones artificiels réalisés à partir de ce type de neurones sont ainsi inspirés du système nerveux. Ils sont conçus pour reproduire certaines caractéristiques des mémoires biologiques par le fait qu'ils sont massivement parallèles, capables d'apprendre et de mémoriser l'information dans les connexions entres neurones, capables de traiter des informations incomplètes. En 1949, Hebb a mis en évidence l'importance du couplage synaptique dans l'apprentissage par renforcement ou dégénérescence des liaisons inter-neuronales lors de l'interaction du cerveau avec le milieu extérieur. Le premier modèle opérationnel est le perceptron simple inspiré du modèle visuel et capable d'apprentissage. Il a été proposé en 1958 par Rosenblatt. Les limites du Perceptron monocouche du point de vue performance ont été montrées en 1969 par les mathématiciens Minsky et Papert. Les travaux de Hopfield en 1982 ont montré que des réseaux de neurones artificiels étaient capables de résoudre des problèmes d'optimisation et ceux de Kohonen (1982) ont montré qu'ils étaient capables des résoudre des tâches de classification et de reconnaissance [75].

III.4.2. Neurone biologique

Le neurone est une cellule composée d'un corps cellulaire et d'un noyau. Le corps cellulaire se ramifie pour former ce que l'on nomme les dendrites, celles-ci sont parfois assez nombreuses que l'on parle alors de chevelure dendritique ou d'arborisation dendritique. C'est par les dendrites que l'information est acheminée de l'extérieur vers le soma (corps du neurone). L'information traitée par le neurone est propagée ensuite le long de l'axone (unique) pour être transmise aux autres neurones. La transmission entre deux neurones n'est pas directe. En fait, il existe un espace inter-cellulaire de quelques dizaines d'angströms (10^{-9}m) entre l"axone du

neurone afférent et les dendrites (on dit une dendrite) du neurone différent. La jonction entre deux neurones est appelée la synapse [75] (voir la figure III.5).

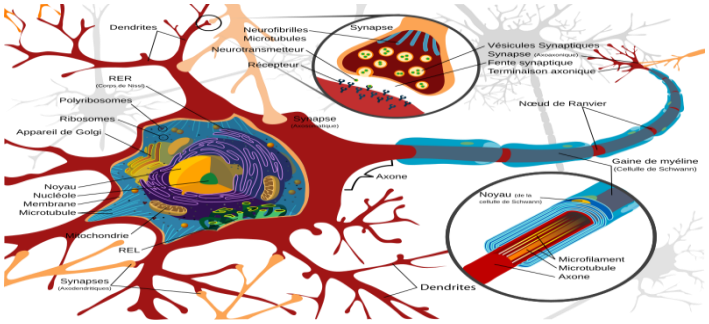


Figure III.5 : Structure d'un neurone biologique [75].

- ✓ **Le corps cellulaire :** C'est l'élément qui contient le noyau du neurone et l'organe biochimique nécessaire pour la production des enzymes, il est de forme pyramidale ou sphérique sa fonction est de faire la somme des influx entrant et envoie lui-même un influx lorsque la somme dépasse un seuil précis via l'axone.
- ✓ **Les dendrites** : Ils se prolongent du corps cellulaire, et ils se ramifient autour du neurone et forment une sorte de vaste arborescence leur fonction est de capter les signaux envoyés au neurone.
- ✓ **L'axone :** Il conduit les signaux émis par le corps cellulaire, sa taille peut varier de quelques millimètres jusqu' à plus d'un mètre, il peut se ramifier à son extrémité comme il peut se connecter aux dendrites des autres neurones.
- ✓ **Les synapses** : Ils sont les points de communication entre les neurones, en général entre l axone d'un neurone émetteur et les dendrites d'un neurone récepteur.

III.4.3. Le fonctionnement du neurone

Le processus de fonctionnement d'un neurone implique la transmission de signaux électrochimiques entre ses différentes parties. Lorsque ces signaux électriques atteignent le bouton terminal du neurone, cela déclenche la libération de neurotransmetteurs qui se déplacent à travers l'espace synaptique et activent les récepteurs du neurone suivant, créant ainsi une série de signaux électrochimiques. Plus précisément, le fonctionnement d'un neurone peut être divisé en trois étapes principales : la transmission de l'influx nerveux, la transmission synaptique et l'intégration des signaux [75].

Le début de la transmission de l'information nerveuse consiste en un signal électrique nommé potentiel d'action, qui se propage le long de l'axone du neurone. Ce processus est permis grâce à une différence de charge électrique entre l'intérieur et l'extérieur de la cellule nerveuse, ainsi que grâce à la présence de canaux ioniques particuliers qui permettent aux ions de traverser la membrane cellulaire.

La deuxième étape du processus est la transmission synaptique, où les neurotransmetteurs sont libérés dans l'espace synaptique entre deux neurones et se lient aux récepteurs de la membrane du neurone suivant, déclenchant ainsi un potentiel d'action.

La troisième étape est l'intégration des signaux, où le neurone reçoit et fusionne les différents signaux électrochimiques provenant de différents neurones, avant de prendre une décision quant à la transmission ou non d'un potentiel d'action.

III.4.4. Neurone artificiel

Appelé aussi neurone formel est un modèle mathématique inspiré d'un neurone biologique (voir la figure III.6) , le neurone artificiel possède plusieurs entrées qui correspond aux dendrites où il recueille l'information et une sortie qui correspond au cône d'émergence (point de départ de l'axone) d'où il envoie un signal électrique, Les actions excitatrices et inhibitrices des synapses sont représentées, la plupart du temps, par des coefficients numériques (les poids synaptiques) associés aux entrées. Les valeurs numériques de ces coefficients sont ajustées dans une phase d'apprentissage (voir le tableau III.1). Dans sa version la plus simple, un neurone formel calcule la somme pondérée des entrées reçues et ajoute un biais, puis applique à cette valeur une fonction d'activation, généralement non linéaire. La valeur finale obtenue est la sortie du neurone, comme montré dans l'équation suivante [75] :

$$y_i = \varphi(\sum_{i=1}^{n}(w_{ij} * x_i) + b_j)$$

- x_i**:** les entrées.
- w_{ij}**:** les poids synaptiques.
- b_j**:** le biais (seuil).
- $\sum_{i=1}^{n}(w_{ij} * x_i) + b_j$**:** la somme pondérée.
- φ**:** la fonction d'activation (sigmoïde, tanh,
- y_i**:** la sortie du neurone.

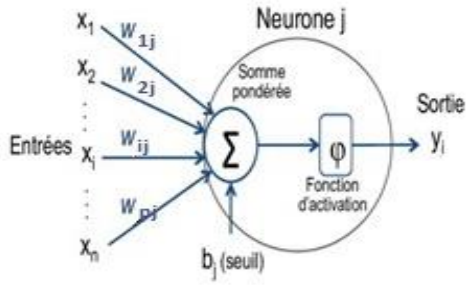


Figure III.6 : Schéma d'un Neurone artificiel [75].

Neurone biologique	Neurone artificiel (formel)
Axones	Signal d'entrée
Synapses	Poids de la connexion
Dendrites	Signal de sortie

Tableau III.1 : Correspondance neurone biologique/neurone artificiel [74].

- ✓ **Le poids :** Le poids est le coefficient qui control le signal d'entrée (la force de connexion), en d'autres termes le poids décide l'influence de l'entrée sur la sortie.
- ✓ **Le biais** : Le biais dans un réseau neuronal est un paramètre supplémentaire qui se connecte aux neurones de la couche précédente via un poids, souvent appelé seuil. C'est similaire à l'ajout d'une constante dans une équation linéaire pour ajuster la sortie en fonction de la somme pondérée des entrées du neurone. En ajustant le biais, le modèle peut mieux s'adapter aux données.
- ✓ **La fonction d'activation** : Est une fonction mathématique responsable du type de l'information émise par le neurone, elle intervient après le calcul de la somme pondérée et l'ajout du biais, elle possède un influe profond sur les performances du réseau. D'où il est préférable de bien choisir le type de cette fonction dans chaque couche du réseau, notons que les neurones d'entrées ne possèdent pas de fonction d'activation, ils utilisent la fonction identité (voir le tableau III.2).

Nom de la fonction	Relation d'entrée/sortie	Icône
seuil	$a = 0$ si $n < 0$ $a = 1$ si $n \geq 0$	
seuil symétrique	$a = -1$ si $n < 0$ $a = 1$ si $n \geq 0$	
linéaire	$a = n$	
linéaire saturée	$a = 0$ si $n < 0$ $a = n$ si $0 \leq n \leq 1$ $a = 1$ si $n > 1$	
linéaire saturée symétrique	$a = -1$ si $n < -1$ $a = n$ si $-1 \leq n \leq 1$ $a = 1$ si $n > 1$	
linéaire positive	$a = 0$ si $n < 0$ $a = n$ si $n \geq 0$	
sigmoïde	$a = \frac{1}{1+\exp^{-n}}$	
tangente hyperbolique	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$	
compétitive	$a = 1$ si n maximum $a = 0$ autrement	C

Tableau III.2 : Différents types de fonctions d'activations [75].

III.4.5. Réseaux de neurones artificiels

Un ANN, également appelé réseau neuronal artificiel, est un système conçu à l'origine pour imiter le fonctionnement des neurones biologiques, mais qui a depuis été développé en utilisant des méthodes statistiques. Les ANNs sont généralement optimisés à l'aide de méthodes d'apprentissage probabilistes, en particulier bayésiennes. Ils font partie de la famille des applications statistiques, apportant des paradigmes permettant des classifications rapides (comme les réseaux de Kohonen), et des méthodes d'intelligence artificielle qui fournissent un mécanisme perceptif indépendant de l'implémentation et des informations d'entrée pour le raisonnement logique formel. En modélisation des circuits biologiques, ils permettent de tester quelques hypothèses fonctionnelles issues de la neurophysiologie, ou encore les conséquences de ces hypothèses pour les comparer au réel [76] (voir la figure III.7**)**.

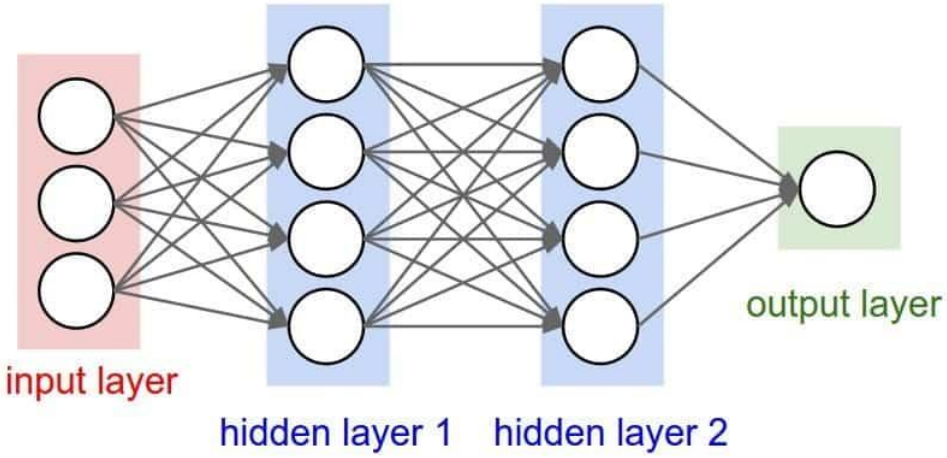


Figure III.7 : Réseau de neurone artificiel [76].

III.4.6. La construction des réseaux de neurones

La construction détaillée d'un réseau de ANN peut varier en fonction de la tâche à accomplir et de l'architecture du réseau choisi. Cependant, voici les étapes générales à suivre [77] :

III.4.6.1. Collecte et préparation des données

Pour commencer l'apprentissage d'un ANN, il est nécessaire de collecter les données requises. Ces données doivent être préparées pour accomplir la tâche souhaitée et doivent être divisées en trois ensembles : l'ensemble d'apprentissage, l'ensemble de validation et l'ensemble de test. L'ensemble d'apprentissage est utilisé pour entraîner le réseau, l'ensemble de validation est

utilisé pour ajuster les hyperparamètres du réseau, tandis que l'ensemble de test est utilisé pour évaluer les performances du réseau.

III.4.6.2. Choix de l'architecture

Il existe différents types d'architectures de ANN, chacune ayant ses propres avantages et inconvénients. Il est crucial de choisir l'architecture la plus appropriée pour la tâche à réaliser. Les architectures populaires comprennent les réseaux de neurones à couches denses, les réseaux de neurones convolutifs, les réseaux de neurones récurrents et les réseaux de neurones avec attention.

III.4.6.3. Définition des couches et des paramètres

Une fois l'architecture choisie, il faut définir les différentes couches du réseau, qui sont des unités de traitement de l'information. Les paramètres de chaque couche sont également définis, comme le nombre de neurones et la fonction d'activation.

III.4.6.4. Définition de la fonction de coût et de l'optimiseur

Pour évaluer la différence entre la sortie prédite par le réseau et la sortie réelle, une fonction de coût est appliquée. Pour réduire cette différence, l'optimiseur est utilisé pour ajuster les paramètres du réseau. L'objectif de l'optimiseur est de minimiser la fonction de coût.

III.4.6.5. Entraînement du réseau

L'entraînement du ANN débute avec l'initialisation aléatoire de ses poids et biais. Les données d'apprentissage sont ensuite utilisées pour ajuster les paramètres du réseau en se basant sur la fonction de coût. Le processus d'entraînement continue jusqu'à ce que la fonction de coût soit réduite au minimum, ou lorsque le réseau ne parvient plus à s'améliorer davantage.

III.4.6.6. Validation et ajustement des hyperparamètres

Une fois l'entraînement terminé, le réseau est évalué sur les données de validation pour déterminer s'il est surajusté (overfitting) ou sous-ajusté (underfitting). Les hyperparamètres du réseau peuvent alors être ajustés pour améliorer les performances.

III.4.6.7. Évaluation des performances

Une fois que les hyperparamètres ont été ajustés, le réseau est évalué sur les données de test pour déterminer ses performances sur des données non vues auparavant.

III.4.7. Architecture des réseaux de neurones

On distingue deux structures de réseau, en fonction du graphe de leurs connexions, c'est-à-dire du graphe dont les nœuds sont les neurones et les arêtes sont les connexions entre ceux-ci [77] :

- Les ANNs statiques (ou acycliques, ou non bouclés).
- Les ANNs dynamiques (ou récurrents, ou bouclés).

III.4.7.1. Les ANNs non bouclés

Un ANN non bouclé est un ensemble de neurones « connectés » entre eux, l'information circulant des entrées vers les sorties sans « retour en arrière ». On peut alors représenter le réseau par un graphe acyclique dont les nœuds sont les neurones et les arêtes sont les connexions entre ceux-ci. Si l'on se déplace dans le réseau, à partir d'un neurone quelconque, en suivant les connexions et en respectant leurs sens, on ne peut pas revenir au neurone de départ. La représentation de la topologie d'un réseau par un graphe est très utile, notamment pour les réseaux bouclés, les neurones qui effectuent le dernier calcul de la composition de fonctions sont les neurones de sortie ; ceux qui effectuent des calculs intermédiaires sont les neurones cachés [77] (voir la figure III.8).

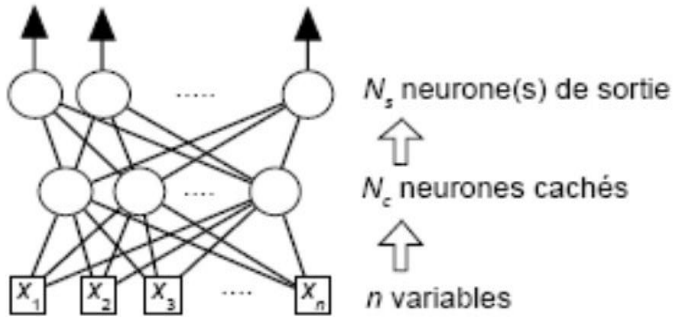


Figure III.8 : Architecture d'un RN non bouclé [78].

III.4.7.2. Les réseaux de neurones bouclés

L'architecture la plus générale pour un ANN est le « réseau bouclé », dont le graphe des connexions est cyclique, lorsqu'on se déplace dans le réseau en suivant le sens des connexions, il est possible de trouver au moins un chemin qui revient à son point de départ (un tel chemin est désigné sous le terme de « cycle »). La sortie d'un neurone du réseau peut donc être fonction d'elle-même, cela n'est évidemment concevable que si la notion de temps est explicitement prise en considération. Ainsi, à chaque connexion d'un réseau de neurones bouclé (ou à chaque arête de son graphe) est attaché, outre un poids comme pour les réseaux non bouclés, un retard, multiple entier (éventuellement nul) de l'unité de temps choisie. Une grandeur, à un instant donné, ne pouvant pas être fonction de sa propre valeur au même instant, tout cycle du graphe du réseau doit avoir un retard non nul. Les connexions récurrentes ramènent l'information en arrière par rapport au sens de propagation défini dans un réseau multicouches. Ces connexions sont le plus souvent locales. Pour éliminer le problème de la détermination de l'état du réseau par bouclage, on introduit sur chaque connexion « en retour » un retard qui permet de conserver le mode de fonctionnement séquentiel du réseau [70] .

III.4.8. Quelques modèles de réseaux de neurones

III.4.8.1. Perceptron simple

Le perceptron est le premier modèle de réseau de neurones inventé en 1957 par Frank Rosenblatt. Le but du perceptron est d'associer des formes en entrée à des réponses. Le perceptron se compose de deux couches : la couche d'entrée et la couche de sortie qui donne la réponse correspondant à la stimulation présente en entrée. Les cellules de la première couche répondent en oui/non. La réponse « oui » correspond à une valeur « 1 » et la réponse « non » correspond à une valeur « 0 » à la sortie du neurone. Les cellules d'entrée sont reliées aux cellules de sortie grâce à des synapses d'intensité variable. L'apprentissage du perceptron s'effectue en modifiant l'intensité de ces synapses. Les cellules de sortie évaluent l'intensité de la stimulation en provenance des cellules de la couche d'entrée en effectuant la somme des intensités des cellules actives [80].

Le perceptron doit trouver l'ensemble des valeurs à donner aux synapses pour que les configurations d'entrée se traduisent par des réponses voulues. Pour cela, on utilise la règle d'apprentissage de WindrowHoff (Correction d'erreur) .

Pour que le perceptron simple (voir la figure III.9) puisse apprendre, il doit pouvoir détecter qu'il a fait une erreur et qu'il aurait dû produire la bonne réponse. C'est pourquoi on parle d'apprentissage supervisé. La règle d'apprentissage est dite locale car chaque cellule de sortie apprend de manière indépendante, elle n'a pas besoin de connaître la réponse des autres cellules. En effet, une cellule ne modifie l'intensité de ses synapses (c'est-à-dire qu'elle n'apprend) que lorsqu'elle se trompe.

Marvin Lee Minsky a montré qu'une forme toute simple (le XOR) ne peut être apprise par un neurone de type perceptron. Un neurone ne peut séparer que deux régions séparables par un hyper plan, avec plusieurs neurones, ça va déjà mieux mais il est vite clair qu'une seule couche de perceptron ne peut pas apprendre des figures complexes (voir la figure III.10).

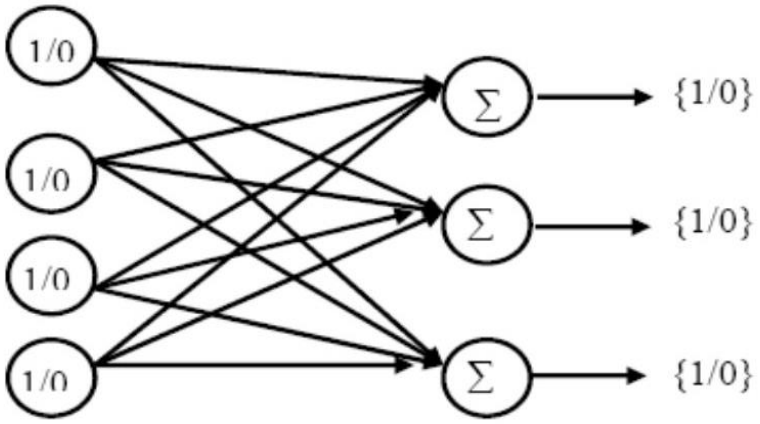


Figure III.9 : Schéma général de perceptron simple [80].

III.4.8.2. Réseaux auto-organisateur (cartes de Kohonen)

Un réseau auto-organisateur, également connu sous le nom de réseau de Kohonen, constitue une catégorie de ANN visant principalement les usages de classification non supervisée et de visualisation de données. Il fonctionne en organisant les données d'entrée sur une grille (souvent en deux dimensions) de neurones, chacun ayant associé un vecteur référent qui caractérise un certain secteur de l'espace des données. Au cours de la phase d'apprentissage, les neurones compétiteurs modifient leurs poids, au profit de ceux qui leur sont voisins, pour arriver à mieux correspondre aux données présentées, permettant ainsi l'élaboration d'une carte topologique capable de préserver les relations des similarités des données d'entrée. Cela permet d'identifier des données similaires sans savoir à l'avance qu'elles le sont, ou du moins, sans

être apte à déduire leur proximité dans l'espace des données à partir d'une variable d'entrée dotée de valeurs des sources d'apprentissage [81] (voir la figure III.10).

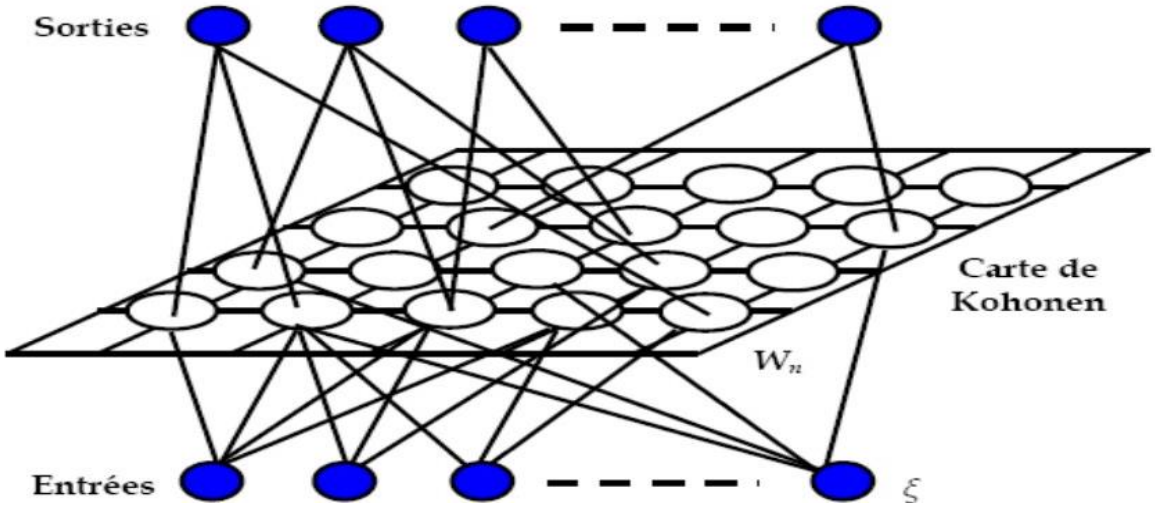


Figure III.10 : Carte topologique auto-adaptative de Kohonen [81].

Les avantages de la carte auto-organisatrice sont :

- L'espace de sortie est un espace de représentation donc on peut visualiser les sorties de la carte.
- Représentation des données de grande dimension.

LUN des inconvénients est le temps de convergence qui est très long. Il n'y a pas de preuve de convergence en multidimensionnel et d'unicité de la représentation.

III.4.8.3. Perceptron multicouche (MLP – Multilayer Perceptron)

Le Multilayer perceptron (MLP) est un type de réseau de ANN largement utilisé dans ML pour effectuer des tâches telles que la classification et la régression. Le MLP est un réseau de neurones qui se compose de plusieurs couches de neurones reliées entre elles. Chaque couche est composée de plusieurs neurones. La première couche est appelée la couche d'entrée, qui reçoit les données en entrée. Les couches intermédiaires sont connues sous le nom de couches cachées, car leurs valeurs ne sont pas directement observées. Enfin, la dernière couche est la couche de sortie, qui génère la sortie finale du réseau. Le MLP utilise une fonction d'activation non linéaire pour chaque neurone. Les fonctions d'activation les plus couramment utilisées sont la fonction sigmoïde, la fonction ReLU (rectified linear unit), la fonction tanh (tangente hyperbolique), et la fonction softmax (pour la classification). Le MLP apprend en ajustant les

poids des connexions entre les neurones, utilisant des algorithmes d'optimisation tels que la descente de gradient. L'objectif est de minimiser la fonction de coût du réseau, qui évalue la différence entre la sortie du réseau et la sortie attendue pour un exemple spécifique. Le MLP est souvent entraîné sur un ensemble de données d'entraînement et testé sur un ensemble de données de test distinct pour évaluer sa performance (voir la figure III.11).

Le MLP a été introduit dans les années 1980 et est toujours largement utilisé aujourd'hui en raison de sa simplicité, de sa capacité à modéliser des relations non linéaires complexes, et de sa bonne performance sur de nombreuses tâches d'apprentissage automatique [82].

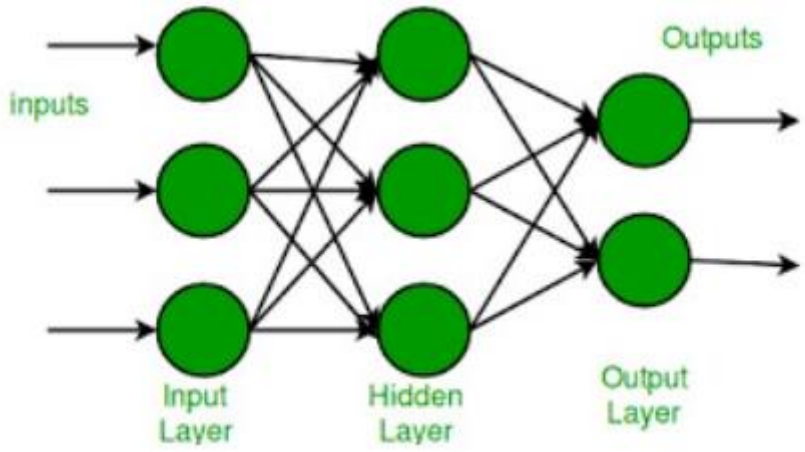


Figure III.11 : Un perceptron multicouche contenant trois couches [82].

III.4.9. L'apprentissage des réseaux MLP

L'apprentissage d'un MLP se fait généralement par l'algorithme de la rétropropagation « back-propagation » qui est l'exemple d'apprentissage supervisé le plus utilisé pour les MLP, Il utilise une méthode d'optimisation universelle consiste à trouver les coefficients du réseau (poids) minimisant une fonction d'erreur globale (fonction coût). La technique de rétro-propagation du gradient est une méthode qui permet de calculer le gradient de l'erreur pour chaque neurone du réseau, de la dernière couche vers la première, le principe de la rétropropagation peut être décrit en trois étapes fondamentales [83] :

1. Initialisation

Avant tout apprentissage, les poids $W^{(l)}$ et les biais $b^{(l)}$ de chaque couche l sont initialisés. En pratique, ils sont choisis avec de petites valeurs aléatoires [83].

2. Propagation avant (Forward Propagation) :

Pour chaque couche l, on calcule :

$$\boldsymbol{z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}} \quad \textbf{Eq III.5}$$

$$\boldsymbol{a^{(l)} = f\left(z^{(l)}\right),\ \ a^{(0)} = x} \quad \textbf{Eq III.6}$$

3. Calcul de l'erreur

La sortie finale $a^{(l)}$ est comparée à la valeur cible y à l'aide d'une fonction de coût, par exemple l'erreur quadratique moyenne (EQM):

$$\boldsymbol{EQM = \frac{1}{2}\left\|y - a^{(L)}\right\|^2} \quad \textbf{Eq III.7}$$

4. Rétropropagation (Backpropagation) :

L'erreur est calculée en sortie puis rétropropagée dans les couches cachées.

Pour la couche de sortie L (y: sortie désirée) :

$$\boldsymbol{\delta^{(L)} = \left(a^{(L)} - y\right).f'^{\left(z^{(L)}\right)}} \quad \textbf{Eq III.8}$$

Pour une couche cachée l :

$$\boldsymbol{\delta^{(l)} = \left((W^{(l+1)})^T \delta^{(l+1)}\right).f'\left(z^{(l)}\right)} \quad \textbf{Eq III.9}$$

5. Mise à jour des paramètres :

Les poids et biais sont corrigés par descente de gradient :

$$\boldsymbol{W^{(l)} = W^{(l)} - \eta \delta^{(l)} (a^{(l-1)})^T} \quad \textbf{Eq III.10}$$

$$\boldsymbol{b^{(l)} = b^{(l)} - \eta \delta^{(l)}} \quad \textbf{Eq III.11}$$

Où η représente le taux d'apprentissage.

6. Répétition :

Les étapes de propagation avant, d'erreur, de rétropropagation et de mise à jour sont répétées pour toutes les données d'entraînement et sur plusieurs epochs, jusqu'à ce que l'erreur EQM soit minimisée.

La phase d'apprentissage d'un réseau de neurone peut donc être résumée par l'algorithme suivant (voir la figure III.12) :

-Initialisation des poids avec des valeurs aléatoires comprises dans un intervalle choisi

-Lecture des exemples d'apprentissage.

-Normaliser les données d'entrainement;

Répéter

Pour chaque exemple d'apprentissage **Faire**

-Propagation de l'entrée vers l'avant

-Propagation de l'erreur vers l'arrière (rétro-propagation)

-Mise à jour des poids

Fin pour

-Calcul de l'erreur totale(EQM)

Tant que l'erreur quadratique moyenne (EQM) est supérieure au SEUIL OU le nombre maximum d'itérations atteint.

Figure III.12 : Algorithme de rétro-propagation de gradient [84].

L'efficacité de l'algorithme de la rétro-propagation dépend, en effet, d'un grand nombre de paramètres que l'utilisateur doit fixer : le pas du gradient, les paramètres des fonctions sigmoïdes, l'architecture du réseau (nombre de couches, nombre de neurones par couche), l'initialisation des poids.

7. Test et évaluation

Une fois le réseau de neurones est entraîné (après l'apprentissage et la validation), il est nécessaire de le tester sur une autre base de données différente de celle utilisée pour l'apprentissage qui est appelée généralement base de test. Ce test permet à la fois d'apprécier les performances du système neuronal en calculant les différentes métriques d'évaluation, telles que la précision, le rappel, le taux de réussite, etc. comme nous illustrons dans le prochain chapitre.

III.4.10. Avantages et Limites des réseaux de neurones

❖ **Avantages :**

- ✓ **Grande flexibilité :** aucune connaissance antérieure des données n'est nécessaire, on les utilise sur différents types de problèmes (image, langage, séries temporelles) [85].

- ✓ **Grande capacité d'apprentissage de représentations hiérarchiques complexes,** ce qui permet d'apprendre à reconnaître des motifs simples puis à les combiner en concepts plus élaborés.
- ✓ **Exécutant de façon efficace sur des données massives et non structurées,** ses performances sont d'un niveau proche de celles des performances humaines.
- ✓ **Résilience et tolérance aux pannes :** le réseau reste fonctionnel même en cas de défaillance de certains neurones.
- ✓ **Les applications sont nombreuses et variés : détection** de fraude, reconnaissance vocale et faciale, diagnostic médical, prévisions financières, robotique, marketing.

❖ **Limites**

- ✓ **Un impératif d'un grand volume de données** de haute qualité pour un apprentissage efficace [85].
- ✓ **Complexité dans la conception**, le choix de l'architecture et l'optimisation de ses paramètres, reposant sur une expertise et des capacités de calcul.
- ✓ **Absence d'interprétabilité des décisions prises,** source de difficultés dans des domaines-sensibles (santé, finance).
- ✓ **Risque d'overfitting**, préjugeant de la capacité de généralisation.
- ✓ **Temps d'entraînement** parfois long et nécessitant de grandes ressources informatiques.

III.4.11. Domaines d'applications des réseaux de neurones

- ➢ Traitement d'image : compression d'images, reconnaissance de caractères et de signatures, reconnaissance de formes et de motifs, cryptage, classification, etc.
- ➢ Traitement du signal : traitement de la parole, identification de sources, filtrage, classification [86].
- ➢ Traitement automatique des langues : segmentation en mots, représentation sémantique des mots, étiquetage morphosyntaxique, traduction automatique, etc.
- ➢ Contrôle : diagnostic de pannes, commande de processus, contrôle qualité, robotique.
- ➢ Optimisation : allocation de ressources, planification, régulation de trafic, gestion, finance, etc.
- ➢ Simulation : simulation boîte noire, prévisions météorologiques.
- ➢ Classification d'espèces animales étant donnée une analyse ADN.
- ➢ Modélisation de l'apprentissage et perfectionnement des méthodes de l'enseignement.

- Approximation d'une fonction inconnue ou modélisation d'une fonction connue mais complexe à calculer avec précision.

III.5. Conclusion

Ce chapitre a permis d'introduire les principes fondamentaux de la classification automatique, une méthode essentielle dans de nombreux domaines tels que la fouille de données, le diagnostic médical ou encore l'intelligence artificielle. Nous avons également rappelé les notions de base ainsi que la typologie des principales techniques de classification, constituant ainsi le socle théorique nécessaire à la mise en œuvre et à l'évaluation des algorithmes présentés par la suite.

Nous avons explicité le fonctionnement de deux classifieurs apprenants supervisés, le KNN et l'ANN, en rendant précises leurs modalités de fonctionnement, leurs principes, leurs clés et leurs points de fonctionnement. Par exemple, l'algorithme KNN utilise les distances entre les points pour attribuer une classe aux entrées, l'ANN notamment à travers le perceptron multicouche s'adapte à travers des itérations à ajuster ses poids, afin d'optimiser le minimum de l'erreur. Ces deux voies exemplifient la large gamme de possibilités d'approche de classification, démontrant bien comment elles seront adaptées à des données médicales.

La répartition en classes représente un maillon essentiel du procédé d'analyse automatique des images médicales, cette étape étant décisive pour détecter des anomalies diagnostiquées. L'utilisation du classifieur le mieux approprié, correspondant à la nature des données et des objectifs d'analyse, est déterminante pour garantir des performances satisfaisantes. Ce chapitre se déclare d'emblée comme l'amorce d'investigations futures pour étendre l'étude aux autres classifieurs, pour espérer d'autant de meilleures valeurs pour le diagnostic assisté par ordinateur. Ainsi, la bonne compréhension des mécanismes régissant les classifieurs étudiés ici est un atout inestimable pour le reste de nos travaux.

CHAPITRE IV :

IMPLÉMENTATION ET ÉVALUATION DE KNN ET ANN SUR LA BASE DDSM POUR LA CLASSIFICATION DES TUMEURS DU SEIN

IV.1 Introduction

Le cancer du sein est l'un des cancers les plus fréquents chez les femmes à l'échelle mondiale. Un dépistage précoce et une classification précise des tumeurs (bénignes ou malignes) jouent un rôle crucial dans le pronostic et le traitement efficace de la maladie. Face à l'augmentation du volume des données médicales et à la complexité des diagnostics, les systèmes intelligents basés sur l'apprentissage automatique offrent des solutions prometteuses pour assister les professionnels de santé dans la détection automatique et la classification des cas suspects.

Dans cette perspective, ce chapitre vise à implémenter et à analyser deux approches de classification supervisée, à savoir le réseau de neurones artificiels (ANN) et l'algorithme des k plus proches voisins (KNN), dans le contexte de la classification du cancer du sein. Les modèles seront entraînés et testés sur la base de données CBIS-DDSM, après avoir fait l'objet d'un prétraitement adéquat. Une évaluation comparative sera ensuite menée en s'appuyant sur plusieurs indicateurs de performance, afin de déterminer la méthode la plus efficace. L'ensemble du processus méthodologique suivi est présenté dans l'organigramme de (voir la figure IV.1).

IV.2 Base de données utilisée et extraction des descripteurs

La base de données utilisée dans cette étude est CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography) qui est disponible sur Kaggle. Il s'agit d'une version enrichie et réorganisée de la base DDSM, contenant des images mammographiques annotées, associées à des métadonnées cliniques et à des masques de segmentation des lésions. Elle fournit, pour chaque image, la localisation des anomalies (tumeurs bénignes ou malignes) sous forme de masques binaires. (Voir la figure IV.2) présente quelques exemples typiques d'images mammographiques issues de cette base, illustrant la complexité et les défis inhérents à la détection et classification des lésions [87].

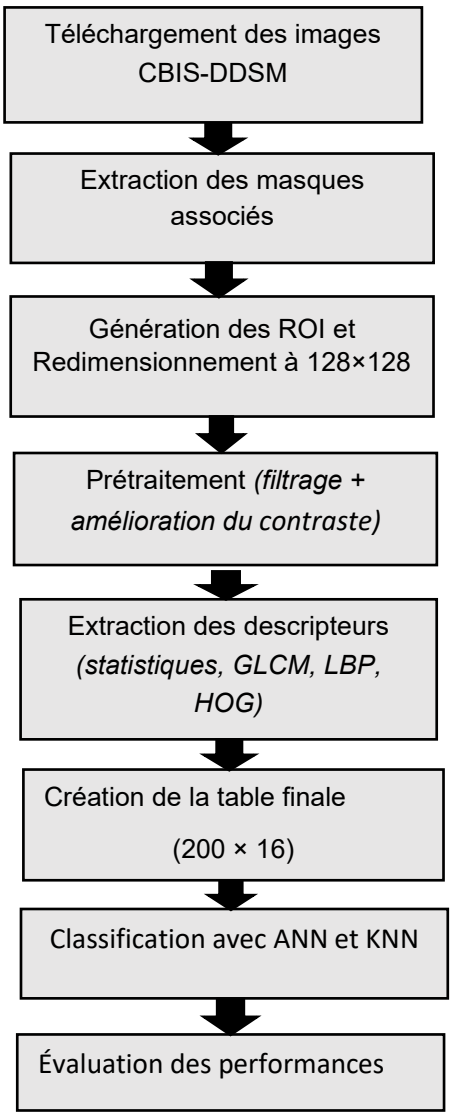


Figure IV.1 : Organigramme du processus de traitement et de classification des images CBIS-DDSM.

4.2.1 Prétraitement et extraction des régions d'intérêt (ROI)

Les régions d'intérêt (ROI) ont été extraites à partir des images d'origine en utilisant les masques fournis avec la base (voir la figure IV.3). Le processus d'extraction s'est déroulé en plusieurs étapes :

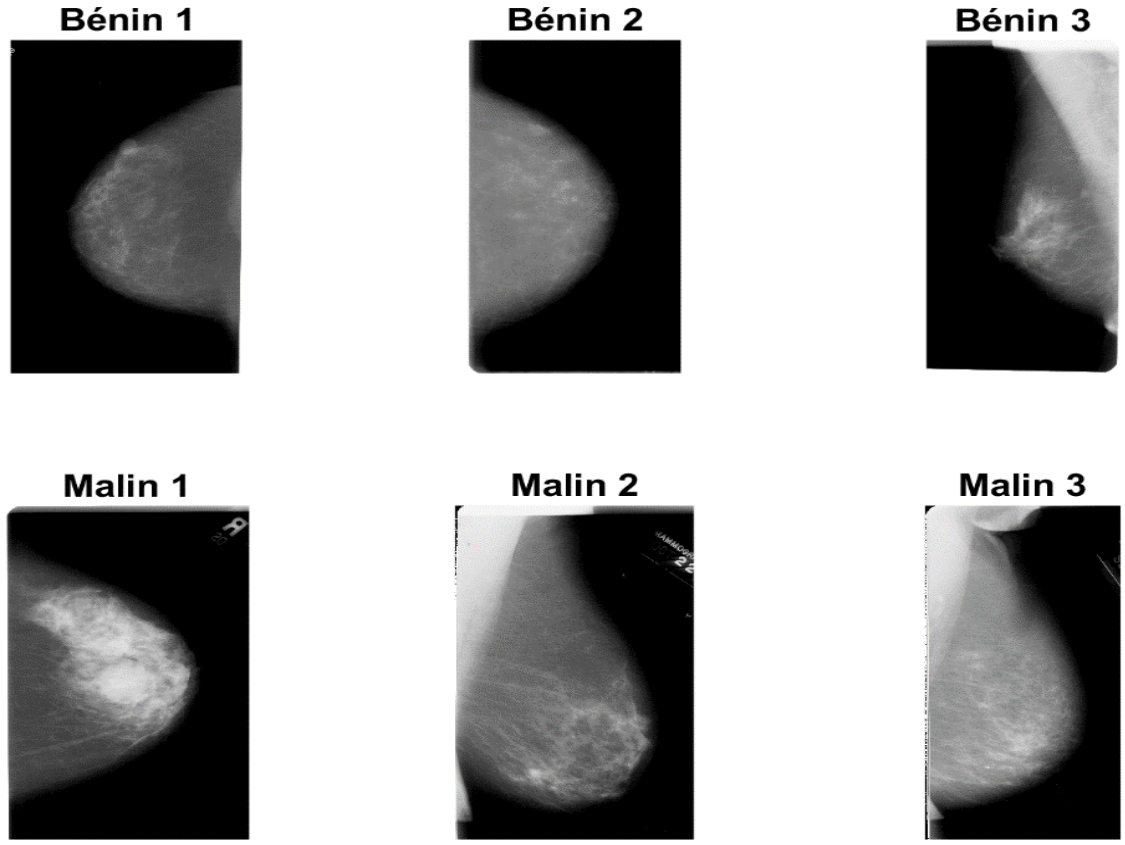


Figure IV.2 : Exemples d'images mammographiques DICOM issues de la base CBIS-DDSM**:**
La ligne supérieure montre trois cas de lésions bénignes, tandis que la ligne inférieure montre trois cas de lésions malignes.

1. Chargement de chaque image mammographique et du masque correspondant.
2. Application du masque sur l'image pour isoler uniquement la zone de la lésion.
3. Découpage du ROI à partir de la zone segmentée.

Une fois extraites, les ROIs ont été redimensionnées à une taille standard de 128×128 pixels afin d'uniformiser l'ensemble des données en entrée (voir la figure IV.4). Cette étape de normalisation permet également de simplifier le traitement et l'entraînement des modèles.

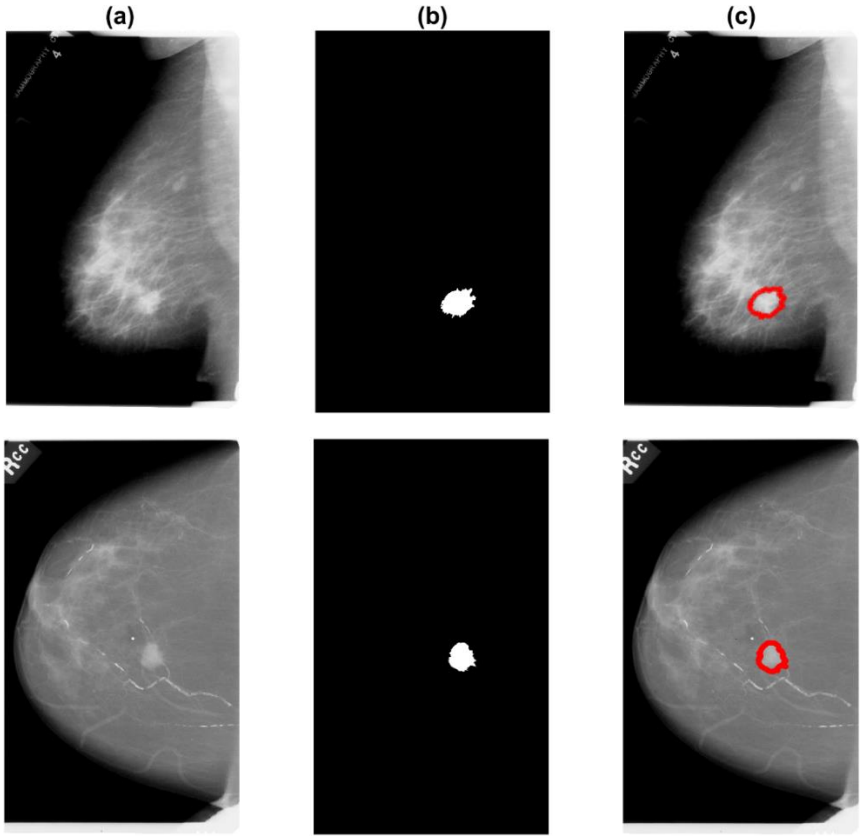


Figure IV.3 : Exemple d'images de mammographie : (a) Images originales, (b) Masques de la lésion, (c) Sélection de la zone d'intérêt.

Par la suite, un prétraitement a été appliqué à chaque ROI pour améliorer la qualité visuelle et l'extraction des caractéristiques. Ce prétraitement inclut un filtrage pour la réduction du bruit (filtrage médian), ainsi qu'un rehaussement de contraste (à l'aide de l'égalisation d'histogramme). Ces traitements visent à accentuer les structures locales utiles à la classification comme la montre (voir la figure IV.5).

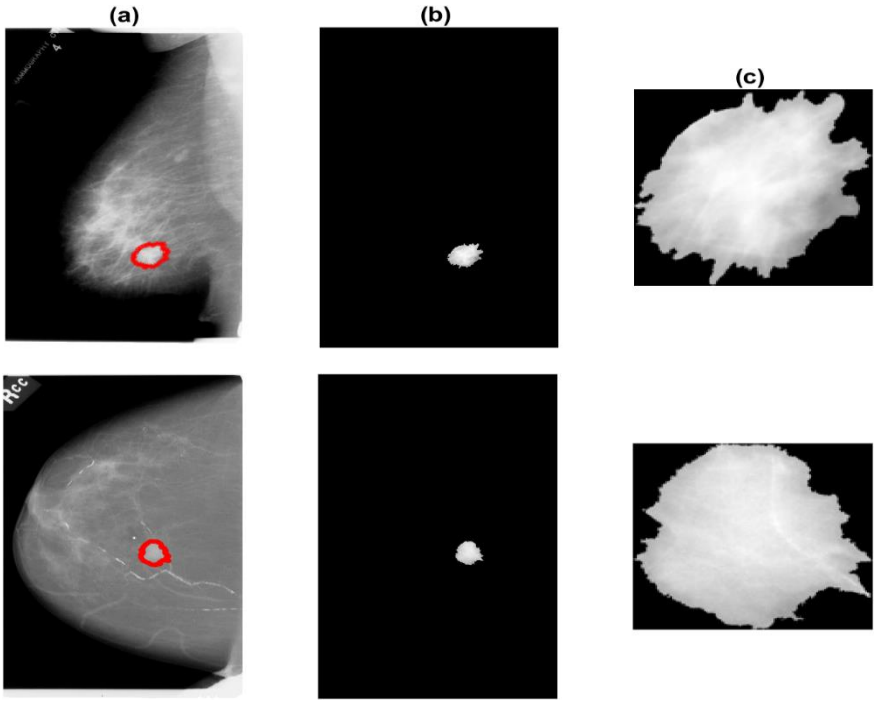


Figure IV.4 : Étapes d'extraction de la lésion : (a) Images originales avec masque superposé, (b) ROI extraites, (c) ROI redimensionnées (128×128).

IV.2.2 Extraction des descripteurs

À partir de chaque ROI prétraité, plusieurs types de descripteurs ont été extraits pour caractériser la texture, la forme et les propriétés statistiques de l'image. Les descripteurs extraits peuvent être regroupés en trois grandes familles :

- **Descripteurs statistiques** : moyenne, écart type, variance, asymétrie (skewness), aplatissement (kurtosis), énergie et entropie. Ces mesures permettent de quantifier la distribution des niveaux de gris dans la région analysée. Ces mesures sont largement utilisées dans l'analyse d'images médicales pour quantifier la distribution d'intensité des pixels [87].
- **Descripteurs texturaux :** basés sur la matrice de co-occurrence des niveaux de gris (GLCM), ces descripteurs sont obtenus à partir d'une matrice de co-occurrence calculée pour différentes orientations (0°, 45°, 90° et 135°) et distances. Les mesures extraites comprennent le contraste, la corrélation, l'homogénéité, l'énergie et la dissimilarité [87].
- **Descripteurs locaux de texture (LBP)** : les motifs LBP permettent de décrire efficacement les structures locales de texture en comparant chaque pixel à son

voisinage. Cette méthode est robuste aux variations d'éclairage et a montré son efficacité dans diverses applications de classification d'images biomédicales [88].

- **Histogramme des gradients orientés (HOG)** : initialement développé pour la détection de personnes dans les images naturelles, le descripteur HOG a été adapté avec succès à la détection de formes et de structures dans les images médicales, notamment pour la reconnaissance de lésions et de contours [89].

IV.2.3 Constitution de la table finale

Dans cette étude, un sous-ensemble de 200 images a été sélectionné à partir de la base CBIS-DDSM. Pour chacune de ces images, un vecteur de 15 descripteurs a été construit en combinant les caractéristiques statistiques et texturales. Ainsi, la table finale est composée de 200 lignes (correspondant aux ROI extraites) et 15 colonnes (correspondant aux descripteurs), à laquelle s'ajoute une colonne indiquant la classe de la lésion (bénigne ou maligne). Cette table a constitué l'entrée des classifieurs appliqués par la suite, à savoir les réseaux de neurones artificiels (ANN) et l'algorithme des k plus proches voisins (KNN), comme cela sera détaillé dans les sections suivantes.

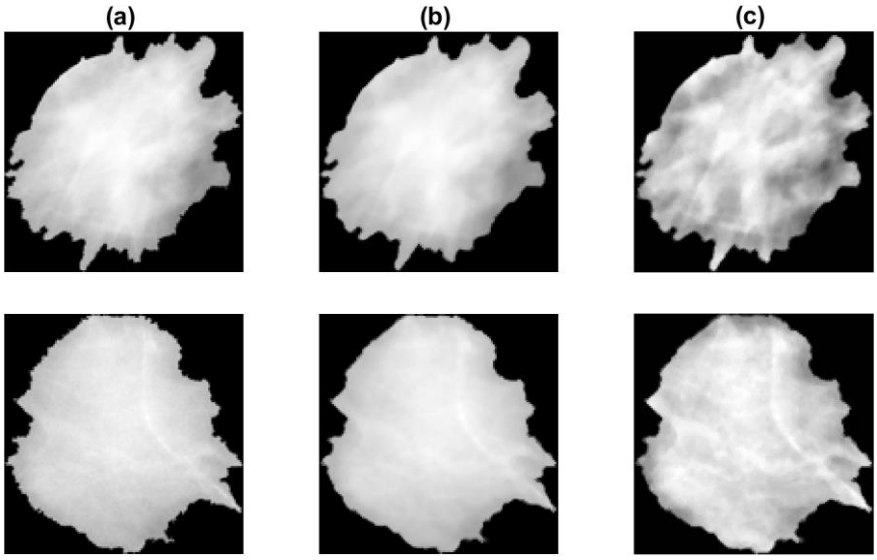


Figure IV.5 : Étapes de prétraitement d'une région d'intérêt (ROI) : (a) images ROI, (b) images filtrées, (c) images filtrées avec amélioration du contraste.

IV.3. Application des Classifieurs KNN et ANN

Cette section présente l'application de deux techniques de classification supervisée : le réseau de neurones artificiels (ANN) et l'algorithme des *k* plus proches voisins (KNN), en vue de

différencier les lésions bénignes et malignes à partir des descripteurs extraits. Toutes les simulations ont été réalisées sous l'environnement MATLAB [88].

IV.3.1 Préparation des données

À partir de la table finale composée de 200 échantillons, chaque observation est représentée par 15 descripteurs extraits à partir des régions d'intérêt (ROI). Ces échantillons sont répartis en deux classes :

- Classe 0 (bénin) : 107 images (soit 53,5 %)
- Classe 1 (malin) : 93 images (soit 46,5 %)

Cette distribution est relativement équilibrée, ce qui permet une évaluation fiable des performances des classifieurs, sans introduire de biais lié à un déséquilibre de classes.

Deux scénarios de division des données ont été testés :

- **Scénario 1** : 70 % pour l'apprentissage (140 échantillons), 30 % pour le test (60 échantillons)
- **Scénario 2** : 80 % pour l'apprentissage (160 échantillons), 20 % pour le test (40 échantillons)

Cette double configuration a permis d'observer l'impact de la répartition des données sur les performances de classification. Avant l'apprentissage, tous les descripteurs ont été normalisés**,** afin de mettre toutes les variables sur une même échelle.

IV.3.2 Classification par KNN

Le classifieur KNN repose sur la mesure de similarité entre un échantillon à classer et ses voisins les plus proches dans l'espace des caractéristiques. Le choix du nombre de voisins k est un paramètre crucial qui influence directement les performances du modèle.

Afin de déterminer la valeur optimale de k, nous avons réalisé une série de tests en faisant varier ce paramètre impair de 1 à 9. Pour chaque valeur, nous avons évalué les performances du classifieur sur notre base de données. Les résultats obtenus (voir la figure IV.6) montrent que la valeur k = 1 donne les meilleures performances en termes de précision globale (Accuracy), ce qui justifie son utilisation dans le reste de cette étude.

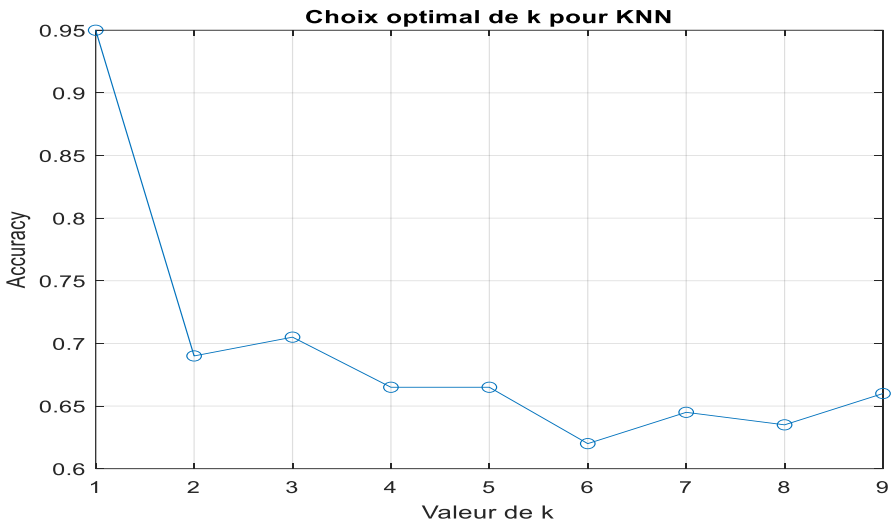


Figure IV.6 : Évolution de la précision du classifieur KNN en fonction du nombre de voisins k pour le scénario 80/20.

IV.3.3 Classification par Réseau de Neurones Artificiels (ANN)

Le classifieur ANN a été développé sous MATLAB à l'aide d'un perceptron multicouche (MLP). Après plusieurs essais empiriques sur l'architecture du réseau, les meilleurs résultats ont été obtenus avec une configuration comportant deux couches cachées **:**

- Couche d'entrée : 15 neurones, correspondant aux 15 descripteurs extraits.
- Première couche cachée : 24 neurones.
- Deuxième couche cachée : 15 neurones.
- Fonctions d'activation : fonctions sigmoïdes utilisées pour les couches cachées, et également pour la couche de sortie.
- Couche de sortie : 1 neurone avec sortie binaire (0 = bénin, 1 = malin).
- Fonction de coût : entropie croisée binaire (binary cross-entropy).
- Optimiseur : rétropropagation du gradient.

Cette architecture a permis d'obtenir une bonne capacité de généralisation, avec des performances supérieures à celles obtenues avec un réseau plus simple à une seule couche.

IV.4 Résultats obtenus

Cette section présente les résultats expérimentaux obtenus des deux classifieurs (KNN et ANN), appliqués sur les deux jeux de données définis précédemment (70/30 et 80/20) [89].

IV.4.1 Évaluation des performances

L'évaluation des performances des deux classifieurs (KNN et ANN) repose sur l'analyse de la matrice de confusion (voir le tableau IV.1), qui permet de comptabiliser les prédictions correctes et incorrectes selon la classe réelle et prédite. Une matrice de confusion présente la répartition des prédictions sous forme de quatre catégories :

	Classe prédite : Positif	Classe prédite : Négatif
Classe réelle : Positif	Vrai Positif (**VP**)	Faux Négatif (**FN**)
Classe réelle : Négatif	Faux Positif (**FP**)	Vrai Négatif (**VN**)

Tableau IV.1 : Matrice de confusion

Les performances ont été évaluées à l'aide des métriques classiques suivantes :

- Exactitude (Accuracy) : proportion globale de prédictions correctes.

$$Accuracy = \frac{VP+VN}{VP+VN+FP+FN} \qquad \textbf{Eq IV.1}$$

- Précision (Precision) : proportion de vrais positifs parmi les positifs prédits.

$$Precision = \frac{VP}{VP+FP} \qquad \textbf{Eq IV.2}$$

- Rappel (Recall ou Sensibilité) : proportion de vrais positifs parmi les cas réellement positifs.

$$Rappel = \frac{VP}{VP+FN} \qquad \textbf{Eq IV.3}$$

- F1-score : moyenne harmonique entre précision et rappel.

$$F1 = 2 \times \frac{Precision \times Rappel}{Precision + Rappel} \qquad \textbf{Eq IV.4}$$

Ces métriques permettent une évaluation plus fine qu'une simple exactitude, en particulier dans le cas de jeux de données déséquilibrés.

IV.4.2 Résultats expérimentaux et matrices de confusion

Les tableaux IV.2 et IV.3 suivants résument les performances obtenues par les deux classifieurs dans les deux scénarios (70/30 et 80/20). Les matrices de confusion associées aux tests en 80/20 sont également présentées pour illustrer concrètement la répartition des prédictions en phase de test (voir les tableaux IV.4 et IV.5).

Classifieur	Répartition	Accuracy (%)	Précision (%)	Rappel (%)	F1-score (%)
KNN (k=1)	70/30	90.5	89.36	90.32	89.84
KNN (k=1)	80/20	95	93.68	95.70	94.68
ANN	70/30	99	97.89	100	98.94
ANN	80/20	100	100	100	100

Tableau IV.2- Bilan des performances de KNN et ANN sur l'ensemble des données.

Classifieur	Répartition	Accuracy (%)	Précision (%)	Rappel (%)	F1-score (%)
KNN (k=1)	70/30	96.67	93.33	100	96.55

Classifieur	Répartition	Accuracy (%)	Précision (%)	Rappel (%)	F1-score (%)
KNN (k=1)	80/20	97.5	95	100	97.44
ANN	70/30	98.33	96.55	100	98.25
ANN	80/20	100	100	100	100

Tableau IV.3 : Bilan des performances de KNN et ANN sur les données de test.

Le classifieur KNN a montré de bonnes performances globales, avec un score de précision et de rappel équilibré. Le scénario 80/20 offre un léger gain d'accuracy, probablement dû à un ensemble d'apprentissage plus riche.

Le réseau de neurones artificiels (ANN) a obtenu des résultats nettement supérieurs à ceux du KNN, en particulier pour le scénario 80/20. L'architecture à deux couches cachées semble fournir une bonne capacité de généralisation. Les résultats obtenus montrent que le réseau de neurones artificiels surpasse le KNN dans toutes les métriques.

	Prédit Malin	Prédit Bénin
Réel Malin	89	4
Réel Bénin	6	101

a. pour l'ensemble des données

	Prédit Malin	Prédit Bénin
Réel Malin	19	0
Réel Bénin	1	20

b. pour les données de test

Tableau IV.4 : Matrice de confusion du classifieur KNN pour le scénario 80/20.

	Prédit Bénin	Prédit Malin
Réel Bénin	93	0
Réel Malin	0	107

a. pour l'ensemble des données

	Prédit Bénin	Prédit Malin
Réel Bénin	19	0
Réel Malin	0	21

b. pour les données de test

Tableau IV. 5 : Matrice de confusion du classifieur ANN pour le scénario 80/20.

Malgré des résultats globalement satisfaisants, les performances du classifieur KNN demeurent inférieures à celles obtenues avec le réseau de neurones artificiels (ANN), notamment en ce qui concerne la robustesse du modèle.

Cela s'explique principalement par le choix du paramètre k = 1. Cette valeur, bien qu'elle ait donné la meilleure précision lors des tests, rend le modèle particulièrement sensible au bruit et aux valeurs aberrantes. En effet, avec un seul voisin pris en compte pour la classification, la décision finale repose sur un échantillon unique, ce qui peut provoquer des erreurs si cet échantillon est mal étiqueté ou atypique.

De plus, le modèle KNN ne réalise aucun apprentissage explicite : il se contente de stocker les données d'entraînement. À l'inverse, l'ANN s'appuie sur un processus d'optimisation (rétropropagation) qui permet d'ajuster les paramètres internes aux données, ce qui lui confère à la fois une capacité d'abstraction, qui lui permet d'extraire des représentations pertinentes à partir des descripteurs, et une capacité de généralisation, qui se traduit par de bonnes performances sur les données de test. Ces deux propriétés expliquent en partie les performances nettement supérieures obtenues avec l'ANN.

IV.5 Conclusion

Dans ce chapitre, nous avons présenté le processus complet de classification des lésions mammaires à partir d'images issues de la base CBIS-DDSM. Après avoir extrait les zones d'intérêt (ROI) à partir des masques fournis, celles-ci ont été redimensionnées, prétraitées, puis décrites par un ensemble de 15 descripteurs statistiques et texturaux (GLCM, LBP, HOG).

Deux classifieurs supervisés ont été implémentés sous MATLAB : le K plus proches voisins (KNN) et le réseau de neurones artificiels (ANN). L'analyse comparative des performances sur

deux répartitions (70/30 et 80/20) montre que l'ANN dépasse significativement le KNN, avec une accuracy atteignant 100 %, contre 97.5 % pour le meilleur cas obtenu avec KNN. Cette différence s'explique en grande partie par la nature même du KNN, qui, avec un k=1, est très sensible au bruit et repose sur des décisions locales. À l'inverse, l'ANN bénéficie d'une phase d'apprentissage qui lui permet de mieux généraliser sur de nouvelles données. Ces résultats confirment que les descripteurs extraits sont discriminants, et que les techniques de classification utilisées permettent une détection efficace des lésions bénignes et malignes. Le réseau de neurones s'est révélé particulièrement performant, ce qui en fait un candidat robuste pour une implémentation dans un système d'aide au diagnostic du cancer du sein.

Conclusion Générale

Le présent mémoire est consacré à la détection et à la classification des lésions mammaires bénignes et malignes à partir d'images mammographiques, en s'appuyant sur des méthodes d'apprentissage automatique, notamment l'algorithme des *k* plus proches voisins (KNN) et les réseaux de neurones artificiels (ANN). Dans un premier temps, nous avons présenté les enjeux médicaux liés au cancer du sein, afin de souligner l'importance du diagnostic précoce et de mettre en évidence les spécificités anatomiques et pathologiques qui conditionnent la conception d'outils de reconnaissance des lésions.

Nous avons ensuite présenté les fondements du machine learning, ses principaux paradigmes ainsi que ses applications majeures, avant de motiver le recours à la classification supervisée à travers les méthodes KNN et ANN. Les zones d'intérêt (ROI) des images issues de la base DDSM ont été extraites à partir des masques fournis, puis redimensionnées et prétraitées. Elles ont ensuite été caractérisées par un ensemble de 16 descripteurs statistiques et texturaux (GLCM, LBP, HOG). Ces étapes de prétraitement et d'extraction des descripteurs se sont révélées cruciales pour garantir la qualité et la pertinence des données mises à disposition des classifieurs.

Les expérimentations menées ont mis en évidence de très bonnes performances pour les deux méthodes testées, avec toutefois une supériorité marquée des réseaux de neurones artificiels, capables d'atteindre une classification parfaite sur l'ensemble des données de test. Cette efficacité s'explique par leur aptitude à modéliser des relations complexes et non linéaires entre les caractéristiques des lésions, faisant des ANN un outil fiable et robuste d'aide au diagnostic assisté par ordinateur.

Ce travail confirme l'importance croissante de l'intelligence artificielle comme soutien aux radiologues, en améliorant la précision, la rapidité et la reproductibilité du diagnostic, tout en réduisant le risque d'erreurs humaines. Cette contribution représente un atout majeur pour la détection précoce du cancer du sein, condition essentielle à de meilleures chances de réussite thérapeutique.

Enfin, cette recherche ouvre des perspectives prometteuses, qu'il s'agisse de l'exploration de techniques d'apprentissage profond plus avancées, de l'intégration d'autres classifieurs (SVM, régression logistique, etc..), de l'optimisation des architectures neuronales ou encore de

l'enrichissement des bases de données. L'utilisation de l'apprentissage automatique s'affirme ainsi comme un levier incontournable pour renforcer la prise en charge du cancer du sein et s'inscrire dans une démarche de médecine plus personnalisée, performante et à forte valeur ajoutée.

Bibliographie

[1] *A. Lecorgne : « Etude épidémiologique, anatomopathologique et immunohistochimique du cancer du sein ». Thèse de doctorat, Université de Bourgogne UFR des sciences de santé circonscription pharmacie, 25-11-2016.*

[2] *[En ligne]. Available: https://ishh.fr/cancer-du-sein/anatomie-et-pathologies-du-sein/ [Accès le 7 septembre 2025]. Publication : 19 février 2025.*

[3] *[En ligne]. Available: https://www.chuv.ch/fr/centredusein/cse-home/patients-et-familles/le-sein-et-ses-maladies/anatomie-du-sein/. [Accès le 7 septembre 2025]. Publication/Mise à jour : 12 juin 2025.*

[4] *I.Hadjij : « Approche morphologique pour la segmentation d'images médicales, application à la détection des lésions mammaires ». Mémoire de magister en électronique biomédicale, Université Aboubekrbelkaid-Telemcen, Faculté des technologies, Département de génie électrique et électronique, 06-07-2011.*

[5] *A. Lecorgne : « Le rôle de pharmacien d'officine dans la prise en charge du cancer du sein après chirurgie mammaire ». Thèse de doctorat, Université de Bourgogne UFR des sciences de santé circonscription pharmacie, 25-11-2016*

[6] *Z. Tahri : « Etude histopathologie et immunohistochimie des cancers mammaires : à-propos de 50 cas ». Mémoire de magister, option cancer et environnement, Université d'Oran-SENIA, faculté des sciences, département de biologie, 2007-2008.*

[7] *Dr N-Belaggoune : « la glande mammaire ». Cours, Université de Batna 2, faculté de médecine, département de médecine ,2020/2021.*

[8] *Korchi-Hammoudi : « Classification des lésions dans les tissus mammaires » Mémoire de master Université Abdelhamid Ibn Badis Mostaganem. 2021/2022*

[9] *[En ligne]. Available: http://www.algeriedz.com/article5733.html. [Accès le 7 septembre 2025].*

[10] *CH. Chiali : « Etude des facteurs de risque du cancer du sein chez des patientes prises en charge au niveau du centre anticancéreux de Sidi Bel Abbes ». Mémoire de master en Biochimie –Immunologie, université Djilali liabes de sidi bel Abbes faculté des sciences de la nature et de la vie département de biologie, 2019/2020*

[11] *M. Julien GAUDAS : « Place de l'estramustine phosphate dans le traitement du cancer du sein métastatique et évaluation de l'expérience du centre expert des maladies du sein de l'hôpital Tenon », Thèse de doctorat en pharmacie, université de paris Descartes, faculté de pharmacie de paris,2016/2017*

[12] *J. Zucman-rossi et J.C. Nault : « Comment une cellule devient cancéreuse et qu'on peut faire pour l'éviter ». The conversation (academicrigourjournalistic flair), 20-12-2016 mise à jour 02-02-2017.*

[13] *[En ligne]. Available: https://www.cancer.fr/personnes-malades/les-cancers/sein/comprendre-les-cancers-du-sein/maladies-du-sein. [Accès le 7 septembre 2025]. Publication mise à jour : 7 juillet 2025.*

[14] *Bart Kova J., Horejsi Z & Koed K., 2005. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. Nature, 434(7035), 864–870*

[15] *F. MOUDJEB et al : « Classification des images de mammographie ». Mémoire de master en électronique biomédicale, UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU Faculté des technologies, Département de génie électrique et électronique, 2016/2017*

[16] *Bendib et al. : « Mémoire de fin d'études sur le cancer du sein en Algérie », Université de Mostaganem, Faculté des Sciences de la Nature et de la Vie, Département de Biologie, 2016.*

[17] *A.Bachir : « Étude épidémiologique du cancer du sein dans la région de Bechar », Université de Bechar, Faculté des Sciences, Département de Biologie, Mémoire de master, année non précisée.*

[18] *A. Yakoub Selloua, Y. Senani : « Cancer du sein chez la femme jeune : Facteurs de risque et prévention » Mémoire de Master, Université 8 Mai 1945 Guelma, Faculté des Sciences de la Nature et de la Vie, Département de Biologie, spécialité Immunologie Appliquée, 2021.*

[19] *A. Abdelouahab, W. Chetibi et al : « Étude épidémiologique et génétique du cancer du sein triple négatif chez la femme algérienne » Université des sciences et de la technologie, Faculté de Médecine, Département de Sénologie, 2019.*

[20] *H. GUENDOUZ, F. ILLIMI, L. RAHAL : « Cancer du sein : Facteurs de risque » Université Alger 1, Faculté de Médecine, Département de Chirurgie Générale, 2024.*

[21] *[En ligne]. Available: https://www.frm.org/fr/maladies/recherches-cancers/cancer-du-sein/focus-cancer-sein. [Accès le 7 septembre 2025]. Publication : 8 juin 2025.*

[22] *A. Thioune, «Décomposition modale empirique et décomposition spectrale intrinsèque : applications en traitement du signal et de l'image,» Université Paris Est Val–de–Marne Créteil, 2016.*

[23] *[En ligne]. Available: https://www.frm.org/fr/maladies/recherches-cancers/cancer-du-sein/focus-cancer-sein. [Accès le 7 septembre 2025]. Publication : 8 juin 2025.*

[24] *[En ligne]. Available: https://chirurgiefemmeparis.fr/cancer-sein/traitement-cancer-sein/carcinome-canalaire-infiltrant/. [Accès le 7 septembre 2025].*

[25] *H . Bouaziz : « Étude de survie du cancer du sein dans la wilaya de Constantine » Université de Constantine 3, Faculté des Sciences, Département de Biologie, 2022.*

[26] *Bendib : « Étude épidémiologique du cancer du sein en Algérie : particularités et prise en charge » Université de Mostaganem, Faculté de Médecine, Département d'Oncologie, 2016.*

[27] *H. BOUAZIZ : « Etude de survie du cancer du sein dans la Wilaya de Ouargla de 2014 à 2019 » Thèse de Doctorat en Sciences Médicales, Université de Constantine 3, Faculté de Médecine, Département de Médecine, 2021.*

[28] *A.Yakoub et al : « Cancer du sein chez la femme jeune : Facteurs de risque et prévention » Mémoire de Master, Université 8 Mai 1945 Guelma, Faculté des Sciences de la Nature et de la Vie, Département de Biologie, spécialité Immunologie Appliquée, 2021.*

[29] *Zergui,mezegeur : « Chapitre I Cancer du sein (diagnostic et prise en charge) » Mémoire, Université Mouloud Mammeri de Tizi-Ouzou, Faculté de Médecine, 2015.*

[30] *Y .Bendjama : « L'imagerie du cancer du sein en Algérie : pratiques et perspectives »Mémoire de master Université d'Alger 1, Faculté de Médecine,Département de Radiologie Année de soutenance : 2021*

[31] *S.Amrani : « Rôle de l'échographie dans le diagnostic du cancer du sein en milieu hospitalier algérien »Thèse de doctorat Université Abou Bekr Belkaid Tlemcen, Faculté de Médecine, Département d'Imagerie Médicale Année de soutenance : 2019.*

[32] *N.Khelifa : « L'impact de la mammogaphie dans le dépistage du cancer du sein en Algérie » Mémoire de master Université de Constantine 3, Faculté de Médecine, Département de Radiologie Année de soutenance : 2022.*

[33] *S.Chabane et N.Haddache : « Segmentation des imagesde mammographies ». Mémoire demaster, Université Abderrahmane mira Bejaia, faculté des sciences exactes, Départementinformatique, 2016-2017.*

[34] *[En ligne]. Available: http://www.breastcare.at/brustkrebsvorsorge/mammographie. [Accès le 8 septembre 2025]. Publication : 31 décembre 2022.*

[35] *M.P.Sampat, M.Markey, A.C.Bovik : « Détection et diagnostic assisté par ordinateur enmammographie ». Manuel de traitement d'image et de vidéos, volume 2, pages 11951217.*

[36] *[En ligne]. Available: http://www.breastcare.at/brustkrebsvorsorge/mammographie. [Accès le 9 septembre 2025]. Publication : 31 décembre 2022 .*

[37] *S.Boudekka : « Prédisposition génétique du gène BRCA1 au cancer du sein » Mémoire, Université Mouloud Mammeri de Tizi-Ouzou, Faculté des Sciences Naturelles et de la Vie, Département de Biologie, 2022.*

[38] *Hadj-Moussa : « Étude sur la mammographie comme outil de dépistage du cancer du sein en Algérie » Mémoire, Université Mentouri Constantine, Faculté de Médecine, Département d'Oncologie,15 mai 2021.*

[39] *A. MEHIDI : « Détection des microcalcifications par les modèles Markoviens pour des tissus mammaires » Thèse de Doctorat en Sciences, Université de Mostaganem, Faculté des Sciences et de la Technologie, Département Électronique, 25 décembre 2019.*

[40] *B. HAOUZI, « Extraction de réseaux linéiques à partir des images à haute résolution » Mémoire en vue de l'obtention du diplôme de magister en Télécommunications et Informatique Spatiales 2012.*

[41] *D. Ikeda, N. Hylton, H. Kinkel et al. « Development, standardization and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance (MR) imaging studies ». J Magn. Reson Imaging 2001 ;13 :896-902.*

[42] *J. Bailar, « Mammography : a contrary view. Annals of Internal Medicine ». 84(1) :77– 84, (1976).*

[43] *K.P Somanand R. Loganathan and V. Ajay. « Machine learning with SVM and other kernel methods ». EasternEconomy Edition, 2009.*

[44] *J. Radiol : « Analyse radiologique des cancers d'intervalle connus, après deux ans decompagne de dépistage de masse organisé (DMO) du cancer du sein en Ille-et- Vilaine ». Edition françaises de radiologie, Paris, vol : 79, pages : 1379-1386, 1998.*

[45] *A. Tradivon :« Quels risques, pour quelles femmes ? Densité mammaire et cancer dusein ». Conférence 30es journées de la SFSPM, LA BAULE, 11-2008.*

[46] *S. AMAROUCHE : « Survie des personnes atteintes de cancer du sein et du cancer colorectal Constantine 2013-2017 » Thèse de Doctorat en Sciences Médicales, Université Salah Boubnider Constantine 3, Faculté de Médecine, Département de Médecine, 2022.*

[47] *F. REGUIED : « Détection semi-automatique de néoplasmes du sein à partir des mammogrammes numériques de la base MIAS » Mémoire de doctorat, École Nationale Polytechnique (ENP) d'Alger, Faculté d'Informatique, Département de Traitement d'Images, 2011.*

[48] *[En ligne]. Available : https://www.cancerimagingarchive.net/collection/cbis-ddsm/. [Accès le 9 septembre 2025]. Publication : 14 septembre 2017.*

[49] *N. DJEFFAL : « Détection assistée par ordinateur pour l'Analyse de mammographies utilisant des techniques d'apprentissage profond » Mémoire de Master, Faculté des Sciences, Département Informatique, Université de Skikda, 2024.*

[50] *M. Benndorf : Digital Database for Screening Mammography (DDSM) « Provision of the DDSM mammography metadata in an accessible format » Mémoire de Doctorat, University Hospital Freiburg, Faculty of Medicine, Department of Radiology, Allemagne, 2014.*

[51] *Wikipédia : « Apprentissage automatique », Mémoire (article en ligne), Fondation Wikimedia, Département d'Intelligence Artificielle, 2025, consulté sur https://fr.wikipedia.org/wiki/Apprentissage_automatique. [Accès le 9 septembre 2025]. Publication de la dernière modification : 9 juin 2025.*

[52] *SAP : ENTREDigital Data base For Screening Mammography (DDSM) « Qu'est-ce que la Machine Learning ? Définition et exemples » MEMOIREOU THESE DE DOCTORAT, 2020.*

[53] *Rahmani A. : « Étude des types d'apprentissage en Machine Learning et leurs applications » Mémoire de fin d'études, Université de Tizi-Ouzou, Faculté des Sciences Informatiques, Département d'Intelligence Artificielle, 2025.*

[54] *F. Ouissal : : « Classification des images mammographiques par deep learning », Mémoire de Master, Université de Guelma, Faculté des Mathématiques, d'Informatique et des Sciences de la matière Département d'Informatique,2022.*

[55] *[En ligne]. Available : https://www.sap.com/france/products/artificial-intelligence/what-is-machine-learning.html. [Accès le 9 septembre 2025]. Publication : 24 mars 2020.*

[56] *C.AFIFI : Deploying Artificial Intelligence techniques for Supporting Decisions in the Business Process Area. Thèse de Doctorat, Université de Guelma, Faculté de Mathématiques et de l'Informatique, Département de l'Informatique, 2025.*

[57] *M.NEMISSI : « Classification et reconnaissance des formes par algorithmes hybrides » Thèse de Doctorat, Université de Guelma, Faculté des Sciences et Sciences de l'Ingénierie, Département de Génie Electrique, 2004.*

[58] *BENALI : « Processus de développement d'un modèle d'apprentissage automatique : de la donnée brute à la mise en production » Mémoire, Université Algérienne, Faculté des Sciences Informatiques, Département d'Intelligence Artificielle, 2025*

[59] *ABBAS J. et HAMDAD A. : « Apprentissage automatique non supervisé pour l'apprentissage du dialecte arabe algérien » MEMOIRE, Université Mouloud Mammeri de Tizi Ouzou, Faculté des Sciences et Technologie, Département d'Informatique, année de soutenance non précisée*

[60] *KHELIFI : « Mise en œuvre d'un processus de clustering automatisé avec vérification manuelle pour la production », Mémoire ou Thèse de Doctorat, Université Algérienne, Faculté des Sciences Informatiques, Département d'Intelligence Artificielle, juin 2025.*

[61] *[En ligne]. Available : https://fr.wikipedia.org/wiki/Apprentissage_semi-supervisé. [Accès le 9 septembre 2025]. Publication : 24 mars 2011.*

[62] *Boumehdi, Nassira, Chekari, Djamila : « Réalisation d'un système d'aide au diagnostic des images mammographiques basé sur le Cloud » MÉMOIRE DE MASTER, Université Ibn Khaldoun -Tiaret-, Faculté des Sciences et Technologie, Département Informatique, 2019.*

[63] *Dupont, Marie :« Apprentissage auto-supervisé pour la classification des images mammographiques sur la base de données DDSM » Mémoire de Master, Université de Paris, Faculté des Sciences et Technologies, Département Informatique, Année de soutenance : 2023.*

[64] *Wikipédia : « Apprentissage_par_renforcement » Mémoire (article en ligne), Fondation Wikimedia, Département d'Intelligence Artificielle, 21 octobre 2006, consulté sur Wikipédia : https://fr.wikipedia.org/wiki/Apprentissage_par_renforcement. [Accès le 9 septembre 2025].*

[65] *Wikipédia : « Apprentissage_par_transfert » Mémoire (article en ligne), Fondation Wikimedia, Département d'Intelligence Artificielle, 12 février 2015, consulté sur Wikipédia [Enligne]. Available : https://fr.wikipedia.org/wiki/Apprentissage_par_transfert. [Accès le 9 septembre 2025].*

[66] *BOUDRAA Sawsen : « Système basé apprentissage et super-résolution pour guider le diagnostic du cancer du sein par mammographie », Thèse de Doctorat, Université Badji Mokhtar Annaba, Faculté de Technologie, Département d'Informatique, 2022.*

[67] *Ahmad, J. : « Deep learning empowered breast cancer diagnosis: challenges and limitations of machine learning approaches using mammographic datasets », Thèse de doctorat,2023*

[68] *Hamzoul, kara : « Application des algorithmes de machine learning pour la prédiction de lien dans les réseaux sociaux » these de doctorat, Université Saad Dahlab, Faculté des Sciences, Département Informatique, 2022.*

[69] *Ultralytics : « Few-Shot, Zero-Shot & Transfer Learning Guide » Blog Ultralytics, 2025. Disponible en ligne sur https://www.ultralytics.com/fr/blog/understanding-few-shot-zero-shot-and-transfer-learning*

[70] *Founder & Chairman @ Junto : « Le Machine Learning : définitions, objectifs et limites » Blog, Junto, 2023.*

[71] *Mohamed Amine Tabouche : « Classification hiérarchique ascendante non supervisée des données » Mémoire, Université de Blida, Faculté des Sciences, Département d'Informatique, 2019.*

[72] *Chahla Benziadi : « Méthodes de Classification » Mémoire, Université Mohamed Khider Biskra, Faculté des Sciences, Département d'Informatique, année non précisée (récente).*

[73] *MOUANE Mohammed Lamine et BENSEDDIK Riad : « L'apprentissage statistique pour le diagnostic de défauts dans un système automatique » Mémoire de Master Professionnel, Domaine : Electronique, Filière : Automatique, Spécialité : Instrumentation et Systèmes, Université Kasdi Merbah Ouargla, Faculté des Nouvelles Technologies de l'Information et de Télécommunications, Département d'Electronique et de Télécommunications, Année 2022.*

[74] ***N.*** *ZOUGHI AGUIR* ***:*** *Université de Carthage, Institut Supérieur des Technologies de l'Information et de la Communication, « Algorithme K-Nearest Neighbors (KNN) : Principes et Étapes », Mémoire, Université de Carthage, Faculté des Sciences et Technologies, Département Informatique, 2021-2022.*

[75] *Moon, Kapila : « Predicting Lung Cancer with K-Nearest Neighbors (KNN) » Mémoire de recherche, Pacific University, Faculté de Sciences, Département d'Informatique, 2024.*

[76] *Cuemath : « Euclidean Distance Formula - Derivation, Examples » Mémoire, Cuemath, Faculté de Mathématiques, Département de Géométrie, 2012.*

[77] *NOUKAS, OUM-HANI : « Etude Comparative des CNNs et de L'algorithme K-NN en mammographie » Mémoire de Master, Université Ibn Khaldoun, Faculté des Sciences et Technologie, Département Informatique, 2023.*

[78] *[En ligne]. Available : https://www.ibm.com/fr-fr/think/topics/knn. [Accès le 9 septembre 2025]. Publication : 3 octobre 2021.*

[79] *Chahla Benziadi : « Méthodes de Classification » Mémoire, Université Mohamed Khider Biskra, Faculté des Sciences, Département d'Informatique, année non précisée ,2019*

[80] *MOUANE BENSEDDIK : « Contribution à l'étude géologique et géotechnique de l'argile plastique du gisement d'El Guerzia (195 km au Sud-Est de Ouargla) en vue de son utilisation en génie civil » Mémoire de Master, Université Kasdi Merbah Ouargla, Faculté des Sciences de la Terre et de l'Univers, Département de Géologie, année de soutenance 2021.*

[81] *AISSOUGUI IHEB : « Modélisation des micromachines/capteurs en utilisant les réseaux de neurones artificiels », Mémoire, Université 8 Mai 1945 Guelma, Faculté des Sciences et de la Technologie, Département d'Électrotechnique, année de soutenance 2019.*

[82] *P. Dupuis : « Étude et implémentation des réseaux de neurones perceptrons multicouches (MLP) pour la classification » Mémoire de Master, Université de Grenoble, Faculté des Sciences et Techniques, Département d'Informatique, 2023.*

[83] *L. Martin : « Analyse de la propagation avant dans les réseaux de neurones artificiels pour la reconnaissance de formes » Thèse de Doctorat, Université de Lyon, Faculté des Sciences, Département d'Informatique, 2021.*

[84] *S. Petit : « Optimisation des réseaux de neurones artificiels via l'algorithme de rétropropagation » Mémoire de Master, Université de Paris, Faculté des Mathématiques et Informatique, Département d'Intelligence Artificielle, 2022.*

[85] *M. Moreau : « Étude critique des réseaux de neurones artificiels : avantages, limites et perspectives » Thèse de Doctorat, Université de Bordeaux, Faculté des Sciences et Technologies, Département de Mathématiques Appliquées, 2024.*

[86] *A. Bernard : « Applications des réseaux de neurones artificiels dans le diagnostic médical et le traitement du signal » Mémoire de Master, Université de Toulouse, Faculté des Sciences de l'Ingénieur, Département d'Informatique, 2023.*

[87] *[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621.*

[88] *[Ojala et al., 2002] Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987*

[89] *[Dalal & Triggs, 2005] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893.*