الجمهورية الديمقراطية الشعبية الجزائرية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

Présenté par

Khaled Khodja Anya

et

Benrached Lilia

pour l'obtention du diplôme de Master en Électronique

Option 'Automatique et Informatique Industrielle'

Thème

Classification morphologique des tumeurs mammaires pour l'aide à la décision

Proposé par : Dr. Reguieg F. Zohra & Pr. Benblidia Nadjia

Année Universitaire 2019-2020

DEDICACES

Je dédie cet humble travail à :

- *Mes chers parents qui se sont sacrifiés pour mon bonheur, que dieu vous protège et vous garde à mes côtés.*
- *A ma chère sœur Neila.*
- *A la mémoire de ma défunte amie d'enfance TAOURIT Yousra décédée en novembre 2019 suite à une longue maladie.*
- *A tous mes amis avec qui j'ai passé d'agréables moments et avec qui cette expérience a été des plus formidables (particulièrement Lilia, Ilham et Bouchra).*

Anya

Dédicaces

Je dédie cet humble travail à :

- *Mes chers parents qui se sont sacrifiés pour mon bonheur, que dieu vous protège et vous garde à mes côtés.*
- *A la mémoire de mon défunt frère BENRACHED CHAKIB décédé en octobre 2018.*
- *A tous mes amis avec qui j'ai passé d'agréables moments, précisément mes trois sœurs… Anya, Hinda et Ilhem.*
- *A mon très cher ami Aymen*

Lilia

Remerciements

Nous remercions d'abord le créateur de l'univers qui nous a doté d'intelligence, et nous a maintenu en santé pour mener à bien cette année d'étude.

Nous souhaitons exprimer un vif remerciement à notre promotrice Madame F. Zohra REGUIEG sans qui cet humble travail n'aurait pu exister, elle a su nous guider et nous faire progresser tout au long de cette recherche et pour cela nous lui sommes reconnaissantes.

Nous tenons aussi à remercier notre co-promotrice Madame Nadjia Benblidia, pour ses encouragements, ses conseils et son soutien, pour notre travail.

Nous tenons également à remercier les Membres du jury, pour l'honneur qu'ils nous font en acceptant de juger ce travail, et de participer à la soutenance.

Nous adressons un grand merci à nos parents. Qu'ils trouvent dans la réalisation de ce travail l'aboutissement de leurs efforts ainsi que l'expression de notre plus affectueuse gratitude.

Nous remercions également nos familles qui nous ont toujours soutenues, un soutien qui a été renforcé par la présence d'amis merveilleux qui nous ont accompagnés pendant notre cursus.

Enfin, nous souhaitons exprimer toute notre reconnaissance à toute personne nous ayant aidé d'une manière ou d'une autre à finaliser ce travail.

A tous, nous exprimons ici notre vive gratitude.

ملخص

يقدم هذا العمل نظام تصنيف مورفولوجي لأورام الثدي من البيانات الخلوية لدعم القرار. في هذا السياق ، ترتدي تقنيات التعلم الآلي والعميقة ؛ من خلال إنشاء مصنف بثلاثة نماذج من خوارزميات التعلم الآلي ونموذج للشبكات العصبية التلافيفية. الهدف هو تقييم أفضلها ، وتكييف معلمات المصنف. النتائج التي تم الحصول عليها تولد دقة 99٪ للشبكات التلافيفية.

الكلمات المفتاحية: أورام الثدي ، التصنيف ، التعلم الآلي ، التعلم العميق ، الشبكات العصبية التلافيفية ، دعم القرار.

Résumé

Ce travail présente un système de classification morphologique des tumeurs mammaires à partir de données cytologiques pour l'aide à la décision. Dans ce cadre, les techniques d'apprentissage automatique et profond, sont usés ; en créant un classifieur avec trois modèles d'algorithmes d'apprentissage automatique et un modèle des réseaux de neurones convolutifs. L'objectif est d'évaluer le meilleur d'entre eux, en adaptant les paramètres du classifieur. Les résultats obtenus, génèrent une précision de 99% pour les réseaux convolutifs.

Mots clés : Tumeurs mammaires, Classification, Apprentissage automatique, Apprentissage profond, Réseaux de neurones convolutifs, Aide à la décision.

Abstract

This work presents a system of morphological classification of mammary tumors from cytopathologic data, for decision support. In this context, machine learning and deep learning, are used; by creating a classifier with three models of machine learning algorithms and a model of convolutional neural network. The aimed objective is to evaluate the best of them by adapting the parameters the classifier settings. The results obtained, generate a 99% accuracy for convoluted networks.

Keywords : Breast tumors, Classification, Machine learning, Deep learning, Convolutional neural networks, Decision support.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

THS : Traitement hormonal substitutif

CAN : Convertisseur analogique numérique

IA : Intelligence Artificielle

K-NN : K-Nearest Neighbors

SVM: Support Vector Machine

MLP: Perceptron multicouche

CNN : Réseaux de neurones convolutionnels

RNN : Réseaux de neurones récurrents

LSTM : Long short terme memory

YOLO: You only look once

Introduction générale

Le cancer du sein est le cancer le plus fréquent et meurtrier chez la femme. Il représente plus du tiers, de l'ensemble des nouveaux cas du cancer chez la femme. Lorsque celui-ci est diagnostiqué à un stade précoce, la patiente guérit dans 9 cas sur 10 [1]. Pour que le taux de mortalité de cancer soit réduit, la tumeur doit être prise en charge dès l'apparition des premiers symptômes (le premier stade). C'est pour cela que le diagnostic est une étape primordiale, dans la lutte du cancer du sein. La mammographie reste la technique par excellence, pour l'exploration du sein et le dépistage précoce de ce cancer. Elle met en évidence les anomalies comme les masses et les calcifications, qui sont ensuite traduites en tumeurs bénignes ou malignes. L'imagerie médicale jour donc un rôle clé. Dans certains cas douteux, l'examen cytologique, un examen complémentaire à la mammographie, est nécessaire pour confirmer le diagnostic.

Dans ce cadre, en cytologie, le cytopathologiste doit établir à l'issu d'examens minutieux de lames de cellules, un diagnostic qui doit être le plus fiable possible. Le pathologiste reconnait les types cellulaires présents sur une lame et, c'est cette étude de la lame, qui détermine le diagnostic. Cet examen se fait via le microscope. Il est donc visuel, manuel et par conséquent très fastidieux et, pourrait engendrer des erreurs de lecture. Afin de faciliter la lecture des lames, des recherches ont permis de développer des systèmes semi-automatiques permettant un diagnostic plus précis.

Le mémoire présenté, s'intègre dans ce contexte dans le but d'une réalisation d'une classification morphologique des tumeurs mammaires, pour l'aide à la décision.

Le travail appréhendé dans ce projet, suscite l'intérêt du laboratoire LATSI (laboratoire du traitement du signal et d'image de l'université Saad Dahlab Blida 1), pour la mise en œuvre d'un système de classification des pathologies mammaires. Cela en se basant sur des propriétés de malignité et de bénignité, des tumeurs du sein provenant d'images cytopathologiques.

Une analyse des caractéristiques les plus utiles, permet de prédire si le cancer et malin ou

bénin et à voir les tendances générales, qui peuvent nous aider dans la sélection du modèle. Le but final est de classer les tumeurs mammaires en des catégories malignes et bénignes, suivant les méthodes d'apprentissage automatique et d'apprentissage profond, en se basant sur les réseaux neuronaux convolutifs.

- **Plan du mémoire**

Afin de mener à bien, notre prélude à la recherche, nous structurons notre mémoire selon quatre chapitres :

- **Le premier chapitre** est consacré au contexte médical.
- **Le second chapitre** parcourt quelques méthodes d'apprentissage automatique (machine Learning) mises en œuvre dans le cadre de ce mémoire.
- **Le troisième chapitre** présente l'apprentissage profond (Deep Learning), ainsi que les modèles des réseaux convolutifs.
- **Le quatrième chapitre** appréhende la méthodologie adoptée, son application ainsi que les résultats obtenus sur des données cytologiques réelles.

Chapitre 1 Contexte Médical

1.1 Introduction

Le cancer du sein et le cancer le plus fréquent chez la femme : une femme sur huit a été, est, ou sera touchée par cette maladie [2]. Chaque année, ce sont près de 49000 femmes pour qui un cancer du sein est détecté et près, de 11900 qui décèdent à cause de cette maladie [1]. Cependant, pour réduire le taux de mortalité et, assurer les soins adéquats, une détection précoce du cancer du sein est primordiale.

La mammographie est la technique la plus efficace pour le dépistage et la surveillance du cancer du sein. Elle permet de visualiser les lésions, au niveau du sein susceptibles d'être des tumeurs bénignes ou malignes. C'est dans ce cadre, que la cytopathologie est utilisée, pour confirmer le cas douteux.

Dans ce chapitre nous allons aborder, d'une manière non exhaustive, l'anatomie du sein, les différentes pathologies mammaires, ainsi l'examen mammographique et la cytopathologie.

1.2 Anatomie du sein

De la puberté à la ménopause les seins évoluent tant par leurs formes, que leurs volumes. Dans un sein se trouve la glande mammaire, entourée de tissus adipeux. Cette glande joue un rôle dans la lactation, lors d'une grossesse ou d'un allaitement [3].

Chaque sein (figure 1.1) contient une glande mammaire (qui contient à son tour quinze à vingt compartiments sépares par des tissus graisseux) et, du tissu de soutien qui contient des vaisseaux, des fibres et de la graisse.

Tous les compartiments de la glande mammaire, sont constitués de lobules et de canaux. Le rôle des lobules est de produire le lait, en période d'allaitement. Puis les canaux transportent ce lait, vers le mamelon.

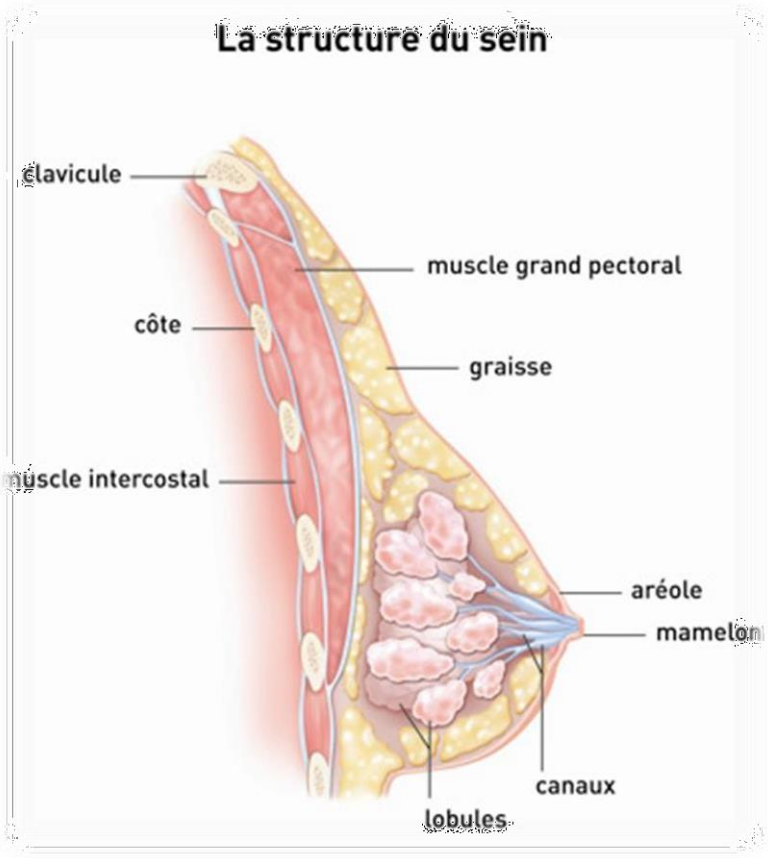


Figure 1.1 : Schéma anatomique du sein [4]

1.3 Cancer du sein

Le cancer du sein (figure 1.2) est une multiplication cellulaire anarchique et incontrôlée, émanant d'une transformation, mutation ou d'une instabilité génétique d'une cellule initialement normale. Cette cellule se multiplie jusqu'à former de grosse masse. Cette masse peut rester localiser dans le sein ou alors de petits fragments de celle-ci peuvent migrais par le biais de la circulation sanguines (vaisseaux sanguins ou lymphatiques) et coloniser d'autres organes : ce sont les localisations à distance de la tumeur ou appelé communément métastase ou bien se localiser uniquement au niveau du sein [5, 6].

La progression d'un cancer du sein prend plusieurs mois voire quelques années.

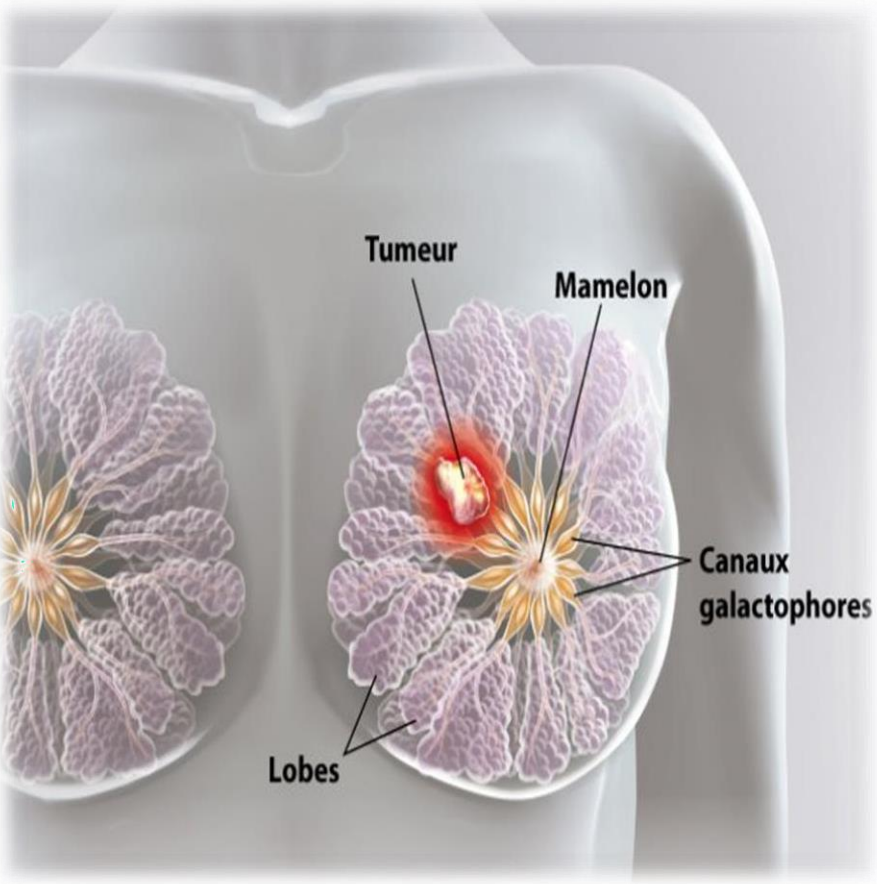


Figure 1.2 : Représentation d'une tumeur mammaire [5]

1.3.1 Symptômes

Il peut n'y avoir aucun signe de symptômes aux premiers stades du cancer. Ces symptômes [6] apparaissent lorsque la tumeur au sein est suffisamment grosse pour être ressentie à la palpation quelques symptômes :

- Une bosse fixe ou mobile au niveau du sein au niveau de l'aisselle.
- Un écoulement provenance du mamelon sans avoir à le comprimé.
- Une rétractation du mamelon (il commence soudainement à pointer vers l'intérieur, mamelon inversé).
- Un changement d'apparence de la peau du sein : un épaississement, un durcissement, une rougeur pas normale, une desquamation autour du mamelon.
- Changement de taille ou de la forme du sein.

1.3.2 Catégories du cancer du sein

Selon la cellule à l'origine du cancer et de l'aspect de la tumeur ou distingue deux types du cancer du sein [7, 8, 9].

1.3.2.1 Cancer non invasif ou non infiltrant

Pour ce cas, on parle d'une tumeur qui ne concerne que les lobules et canaux ; ce qui veut dire que celle-ci ne s'étend pas aux tissus environnants. Dans cette catégorie, on retrouve le carcinome canalaire *in situ,* qui représente le cancer le plus fréquent chez la femme ; se formant ainsi à l'intérieur des canaux de lactation [7]. Le traitement de ce dernier mène presque dans tous les cas, à la guérison. Lors de cas exceptionnels sans traitement, celui-ci termine sa croissance et peur alors devenir infiltrant [8].

Un autre cas de cancer non infiltrant rare, le carcinome lobulaire *in situ* qui lui affecte les lobules [7].

1.3.2.2 Cancer invasif ou infiltrant

Dans ce cas, la maladie envahit les tissus autour des canaux de lactation mais reste à l'intérieur du sein. Lors d'un non traitement, la tumeur peut engendrer des métastases qui vont s'éparpiller dans le corps. Parmi ces cancers :

- ***Le carcinome canalaire***

Ce type de ce cancer se forme dans les canaux de lactation où, les cellules cancéreuses traversent la paroi des canaux de lactation.

- ***Le carcinome inflammatoire***

C'est un cancer rare qui est caractérisé principalement, par un sein qui peut devenir enflé, rouge et chaud. La peau du sein quant à elle, peut avoir une peau d'orange. Ce type de cancer progresse plus vite et il est difficile à traiter.

- ***Le carcinome lobulaire***

Dans ce cas, les cellules cancéreuses apparaissent dans les lobules. Elles traversent ensuite, la paroi des lobules et se propagent dans les tissus environnants.

- ***Le carcinome médullaire***

Il semble bien limité à première vue. Cependant parfois, il se propage aux ganglions axillaires. Les néoplasies de ce genre, peuvent être importantes. Par contre, elles ont un meilleur pronostic que le carcinome canalaire invasif [9].

- ***Le carcinome mucineux ou colloïde***

Ce carcinome est un type de carcinome canalaire invasif, qui est formé d'un nodule de type gélatineux. La néoplasie progresse doucement, mais elle peut devenir très importante à travers le temps. Ce cancer du sein a un très bon pronostic [9].

- ***Le carcinome tubulaire***

Ce cancer produit beaucoup de petites glandes et des tubes, ressemblant énormément aux glandes mammaires et, aux canaux galactophores normaux. Il s'étend souvent aux ganglions axillaires de l'aisselle. Il a cependant, un très bon pronostic [9].

- ***La maladie de Paget***

Elle est rare et se manifeste par une petite plaie au niveau du mamelon, qui ne guérit pas.

1.3.3 Facteurs de risque

Il existe plusieurs facteurs de risque [6, 7] du cancer du sein qu'une personne peut avoir. Cependant, cette personne peut avoir ces facteurs, mais ne jamais avoir de cancer. Voici quelques facteurs :

- **L'Age**

Une personne de plus 50 ans, a plus de chance d'avoir un cancer du sein, qu'une personne de moins de 35 ans.

- **La prédisposition génétique**

Environ 5 à 10% des cancers du sein, sont liés à une prédisposition génétique. Ceci signifie qu'un gène ayant subi des mutations (BRCA1, BRCA2), existe à la base.

- **Les antécédents familiaux**

Ils représentent 20 à 30% du cancer du sein. Le risque de développer un cancer augmente lorsque qu'il a un ou plusieurs membres d'une famille, ayant eux un cancer du sein ou des ovaires.

- **Les antécédents personnels**

Une femme ayant déjà eu un cancer du sein, à un risque 3 à 4 fois plus élevé d'en déclarer un second.

- **La contraception hormonale**

Contraceptif oral contenant des œstrogènes et de la progestérone, elle augmenterait un peu le risque de cancer du sein, lorsqu'elle est prise sur une longue période. Mais dès l'arrêt, ce risque diminue.

- **Le traitement hormonal substitutif (THS) à la ménopause**

Les études ont montré qu'une utilisation prolongée d'un THS, notamment si le traitement contenait à la fois des œstrogènes et de la progestérone, augmenterait légèrement le risque de développer un cancer du sein. Mais le risque semble baisser quelques années, après l'arrêt de celui-ci.

- **Le mode de vie**

La consommation de tabac, d'alcool, ou encore le surpoids et la sédentarité augmenteraient le risque d'avoir un cancer du sein.

1.4 Mammographie

La mammographie est un examen radiologique des seins utilisant des rayons X (figure 1.3). Elle permet le dépistage du cancer du sein ou des lésions précancéreuses (la mammographie de dépistage) ou de diagnostiquer certaines pathologies qui touchent les seins (la mammographie de diagnostic) telles que les nodules, les grosseurs, les douleurs, les modifications de la peau ou des aréoles, des inflammations ou des écoulements du mamelon [10].

1.4.1 Mammographie de dépistage

Cet examen est indiqué chez les femmes entre 50 et 74 ans et doit être réalisé, tous les deux ans ou bien à titre individuel, lorsqu'une femme présente des facteurs de risque particuliers. Ce type d'examen permet de mettre en évidence des cancers de petite taille, à un stade précoce ou avant même l'apparition de symptômes.

1.4.2 Mammographie de diagnostic

Ce type d'examen est effectué lors de la découverte d'un ou plusieurs symptômes dans le sein.

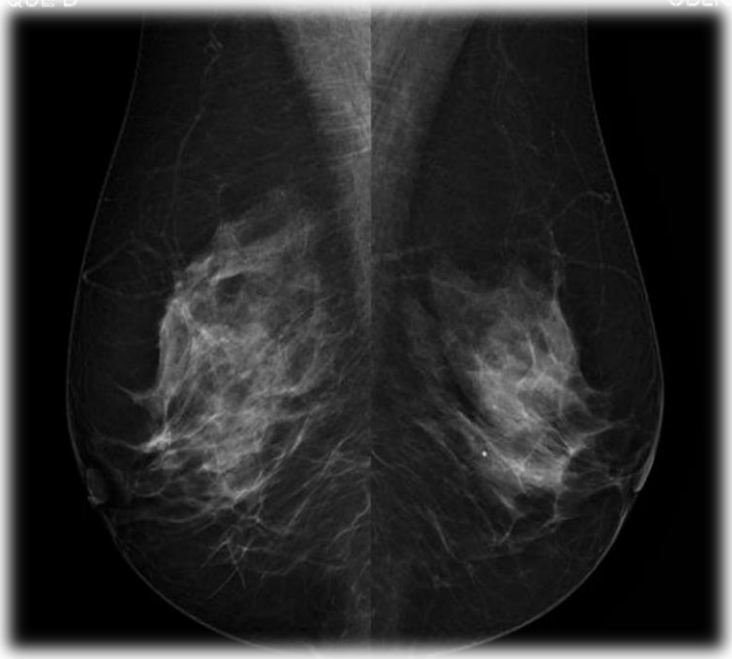


Figure 1.3 : Cliché d'une radio mammographique [10]

1.4.3 Types de mammographie

L'appareil qui réalise les mammographies est appelé mammographe (figure 1.4). Il se compose d'un tube radiogène générateur de rayons x, de faible énergie et d'un système de compression du sein. Cet examen de référence revient à radiographier chaque sein au moins 2 fois (de face et de profil) et parfois même 3 fois (de face, de profil et de biais pour permettre une exploration des creux auxiliaires). Cela se fait en comprimant le sein entre deux plaques de compression, en plastique et un plateau porte-film pour rechercher des anomalies telles que des opacités ou des microcalcifications. Il faut savoir que cette compression doit être suffisante, mais appliquée progressivement pour que la femme puisse bien le tolérer.

La compression se fait pour diminuer l'épaisseur de la glande mammaire, afin de perfectionner le contraste de l'image, d'empêcher le flou lié au mouvement et d'obtenir des clichés de bonne qualité [8, 9, 11].

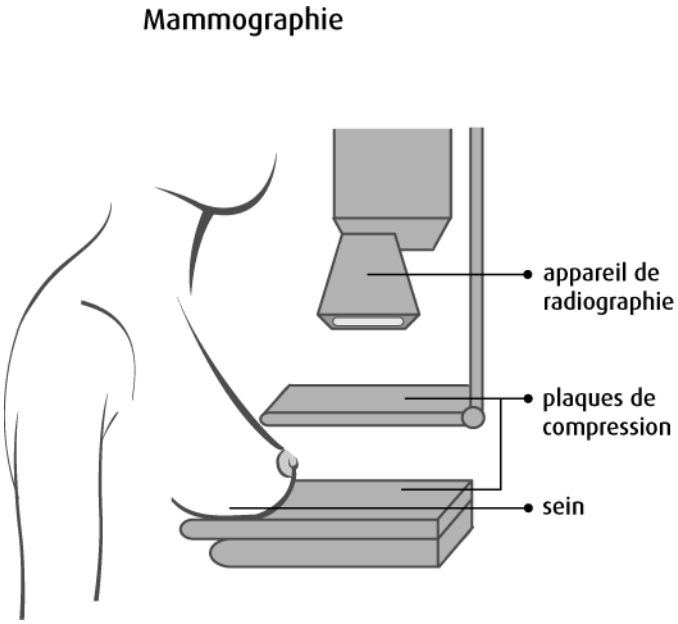


Figure 1.4 : Composantes principales d'un mammographe [11]

1.4.3.1 La mammographie analogique

La mammographie analogique est un système film-écran classique. C'est un film radiographique composé d'un support en polyester, recouvert d'une émulsion de bromure d'argent sur ses 2 faces. Ce film est place dans une 1 cassette ; munie d'un écran. Cet écran placé derrière le film, est réfléchissant à la lumière (photo luminescent) donc il va émettre une lumière lorsqu'il sera exposé au rayons X.

1.4.3.2 La mammographie Numérique

Pour les systèmes numériques, il existe différentes technologies que l'on peut utiliser pour la production d'images, qui se distinguent par le genre de détecteurs utilisé (un détecteur solide ou bien un écran photo luminescent stimulable). Parmi ces systèmes : le système direct, indirect ainsi que la tomosynthèse.

- **Système à conversion indirect**

Les détecteurs de ce système utilisent un scintillateur, qui convertit les rayons X absorbés en photons lumineux, qui sont ensuite transformés en charges électrique, soit par des caméras CCD ou des matrices de photodiodes en silicium amorphe. Le signal obtenu est numérisé, grâce à un convertisseur CAN (Analogique – Numérique).

Les écrans radio luminescents à mémoire, contenus dans une cassette, ont des pièges disponibles pour stocker les électrons créés par ionisation, lors de l'interaction avec les photons X. Ces électrons sont ensuite libérés par photo stimulation, lorsque la surface de cet écran est balayée par un faisceau laser. La libération des électrons s'accompagne d'une émission lumineuse convertie en signal électrique, qui ensuite est numérisée.

- **Système à conversion direct**

Ce système utilise un photoconducteur constitué de sélénium amorphe, comme détecteur. Les rayons X sont directement convertis en charge électrique, afin de produire une image numérique.

- **La tomosynthèse**

La tomosynthèse est une mammographie numérique en 3 dimensions (3D) permettant d'acquérir des images d'un sein compressé, sous différents angles. Tant dis que le sein est immobilisé, le tube à rayons X se déplace sur un arc de cercle au-dessus de lui. Les détecteurs plans sont situés sous le sein. Les coupes mammographiques fines de 1 mm sont envoyées à une console numérique, pour l'analyse et l'interprétation des images. Ce type de technologie permet de diminuer le problème de superposition des tissus, auquel la mammographie classique (2D) est confrontée. Parmi les avantages de la tomosynthèse, une visualisation plus précise de la taille, la forme, la localisation et du nombre d'anomalies détectées dans le tissu mammaire. Il permet aussi une détection précoce, de petits cancers du sein et diminue le taux de rappel [9].

1.5 Pathologies du cancer du sein

La pathologie est la partie de la médecine qui traite de la nature, des causes et des symptômes des maladies, pour au final savoir comment les traiter au mieux et, les prévenir si possible [12, 13].

Il existe différents types de tumeurs qui puissent affecter le sein.

1.5.1 Calcifications mammaires

Ce sont des dépôts de calcium qui se forment, dans le tissu du sein. Elles n'ont aucun lien avec la quantité de calcium absorbée au cours de l'alimentation, ou par l'intermédiaire de complément alimentaires.

Les calcifications mammaires sont assez courantes et sont en général détectées, lors d'une mammographie de dépistage, où elles apparaissent sous forme de points blancs. Le radiologiste observe dans ce cas, la taille, la forme ainsi que la disposition des calcifications.

Notons que la présence de calcifications mammaires ne veut pas dire forcément, cancer du sein. Mais certaines caractérisent de calcifications comme une forme irrégulière ou des regroupements, pour être un signal d'alerte. Il existe deux types de calcifications :

1.5.1.1 Macrocalcifications

Les Macrocalcifications sont des dépôts grossiers de calcium dans le sein, plus fréquent chez les femmes de plus de 50 ans. Elles sont souvent associées à des modifications bégnines, qui se produisent dans le sein et qui sont liées :

- au vieillissement des artères du sein ;
- à d'anciennes lésions ;
- à une inflammation ;
- à des masses telles qu'un fibroadénome.

1.5.1.2 Microcalcifications

Il s'agit de minuscules dépôts de calcium dans le sein. Leur présence signifie parfois, que l'activité de certaines cellules est accrue (une cellule plus active absorbe davantage de calcium que celle qui l'est moins). Un cancer du sein peut être suspecté grâce à elles, surtout lorsqu'elles apparaissent isolées ou regroupées en grappes, à la mammographie. Mais peuvent ne pas être synonymes de cancer.

Si les microcalcifications semblent suspectes, le radiologiste suggère une biopsie, une mammographie de diagnostic, avec compression localisée ou bien une mammographie de contrôle tous les 6 mois.

1.5.1.3 La classification des microcalcifications de Le Gal

Une classification des calcifications est faite selon Le Gal, où on constate 5 types [13].

Type 1 : microcalcifications annulaires rondes à centre clair radio-transparentes. Elles correspondent dans tous les cas, à une pathologie bénigne de galactophorite ectasiante, de microkyste, de liponécose, de dépôts calciques stratifiés circulaires dans l'épaisseur de la paroi de galactophores dilatés.

Type 2 : microcalcifications punctiformes, rondes, pleines, radio-opaques, aux contours réguliers et arrondis. Dans 20% des cas, il s'agit de lésions malignes, dans 20% des cas, il s'agit de lésions frontières, dans 60% des cas, il s'agit de lésions bénignes.

Type 3 : microcalcifications poussiéreuses, trop fines pour préciser leur forme elles donnent une image de semis de poudre calcaire. Des lésions bénignes à 50% et malignes à 50%.

Type 4 : microcalcifications punctiformes irrégulières aux contours anguleux différents d'une microcalcification à l'autre. Dans 70% des cas il s'agit de lésions malignes.

Type 5 : microcalcifications vermiculaires, elles ont la forme d'un bâtonnet souvent irrégulier. 100% Des lésions maligne dans la plupart des cas c'est un comédo-carcinome.

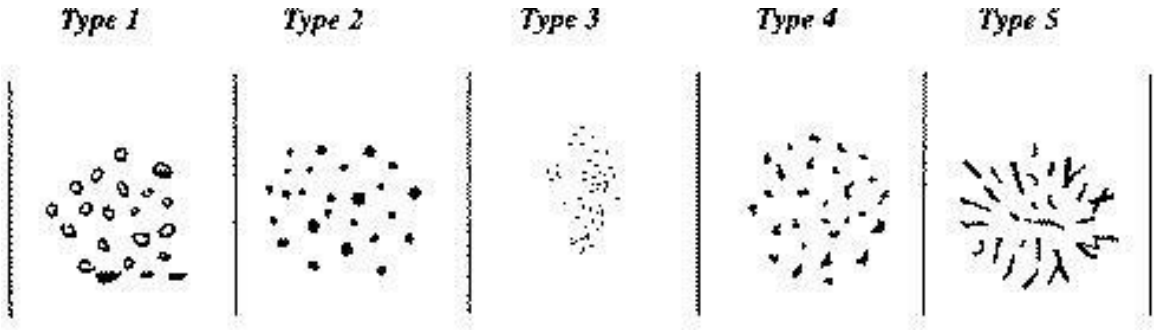


Figure 1.5 : Différentes formes de microcalcifications [12]

Type1 : anneaux, type 2 : punctiformes réguliers, type 3 : poussiéreuses,
Type 4 : punctiformes irrégulières, type 5 : vermiculaires

1.5.2 Les masses

Une masse (opacité), est une lésion occupant une position dans un espace qui peut être vue, sous deux incidences différentes (figure 1.6). Cette opacité peut être difficilement perceptible en mammographie parce qu'elle est similaire aux tissus qui l'entoure et c'est pour cette raison que sa détection peut s'avérer assez difficile et délicate. Sa caractérisation est basée sur sa forme, ses contours et sur sa densité [12]. On distingue plusieurs masses :

a) La forme

Les masses mammaires (figure 1.6) peuvent avoir une forme ovale, ronde, lobulée ou irrégulière.

- **Ronde :** masse circulaire, sphérique ou globuleuse.
- **Ovale :** masse qui a la forme d'un œuf, elliptique.
- **Lobulée :** la forme de la masse présente une ondulation légère.
- **Irrégulière :** c'est des masses dont la forme est aléatoire.

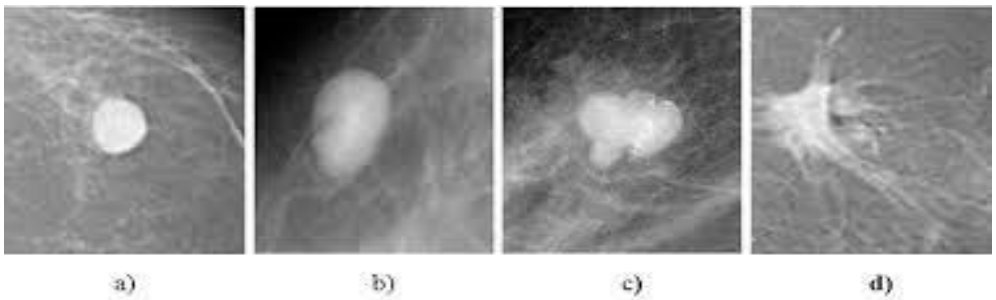


Figure 1.6 : Formes d'une masse [12]

a)Ronde, b) ovale, c) lobulée, d) irregulière

b) Le contour

Le contour des masses mammaires peut être circonscrit, microlobulé, masqué, indistinct ou spiculé.

- **Circonscrit :** c'est une transition brusque, entre la lésion et le tissu environnant, les contours sont nets et viens défini.

Pour dire d'une masse qu'elle est circonscrite il faut qu'au moins 75% de son contour soit délimité.

- **Microlobulé :** le contour crée de petites ondulations, dut à de courtes dentelures.
- **Masqué :** lorsque le contour est caché par le tissu normal adjacent on parle alors de contour masqué.
- **Indistinct :** le contour n'est pas bien défini. Ce caractère indistinct, peut correspondre à une infiltration.
- **Spiculé :** la masse est caractérisée par des lignes radiaires, prenant naissance sur le contour de la masse. Ces lignes sont appelées des spicules.

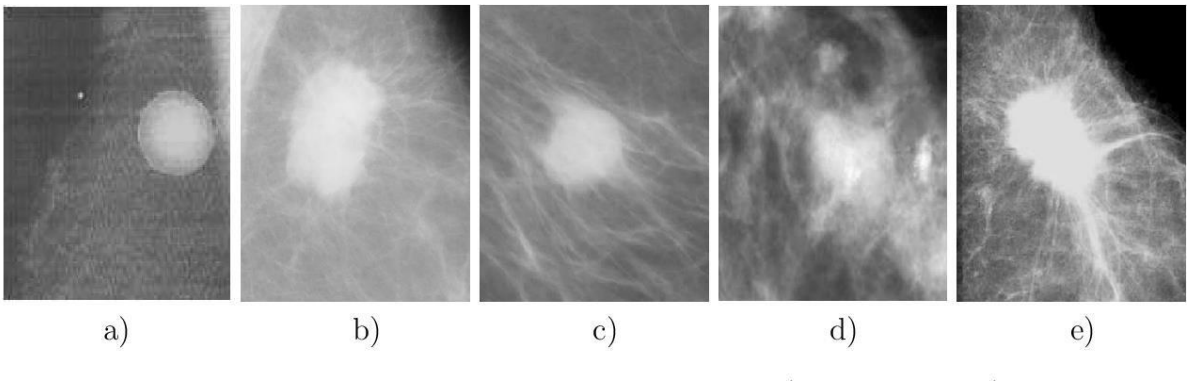


Figure 1.7 : Différents types de contour [12]
a) Circonscrit, b) microlobulé c) masqué, d) spiculé

c) La densité

La densité mammaire (figure 1.8) est mesurée par la proportion du tissu fibro-glandulaire, par rapport au tissus fibreux et donc plus le tissus fibro-glandulaire est important plus le sein est dense.

La classification BIRARDS a défini quatre catégories de quantification de la densité mammaire :

- **Densité type 1 :** les seins sont sombres, presque entièrement graisseux avec un aspect homogène. Ils comptent moins de 25% de tissu fibro-glandulaire et concerne 40% des femmes.
- **Densité type 2 :** De 25% à 50% de tissus fibro-glandulaire, et concerne 25% des femmes.
- **Densité type 3 :** De 50 à 70% de tissus fibro-glandulaire, et concerne 25% des femmes.
- **Densité type 4 :** les seins représentent un aspect très dense, homogène et contiennent plus de 75% de tissu fibro-glandulaire. Concerne 10% des femmes.

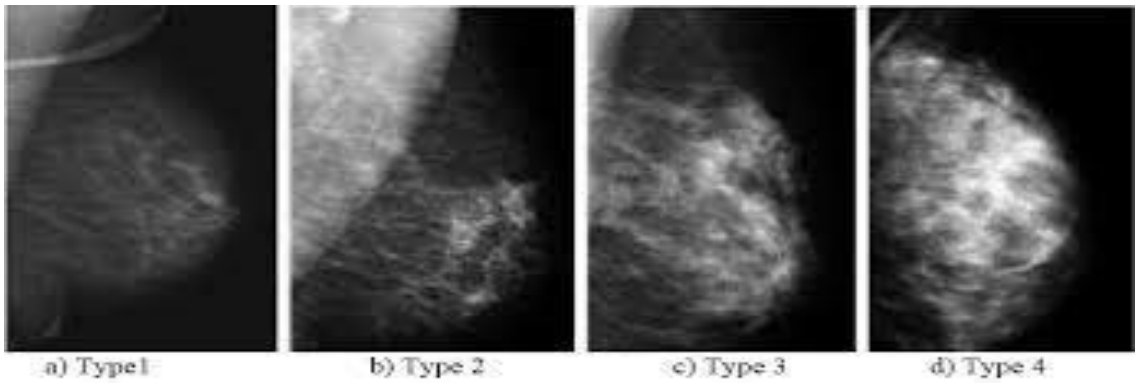


Figure 1.8 : Différents types de densité de masse mammaire [12]

La classification ACR-BIRADS (tableau 1.1) classifie en sept catégories, les images mammographiques et cela en fonction du degré de suspicion de leur caractère pathologique.

Catégorie	Imagerie	Risque de cancer	Conduite
0	Investigation incomplète		Investigation à compléter ou comparer avec films antérieurs
1	Normal	0 %	Retour au dépistage
2	Anomalie bénigne	0 %	Retour au dépistage
3	Anomalie probablement bénigne	> 0 % mais ≤ 2%	Suivi 6 mois
4 4A 4B 4C	Anomalie demandant une biopsie faiblement suspecte modérément suspecte très suspecte	> 2 % mais < 95 % > 2 % à ≤ 10 % > 10 % à ≤ 50 % > 50 % à < 95 %	Biopsie
5	Anomalie fortement suspecte d'un cancer	≥ 95 %	Biopsie
6	Cancer prouvé à la biopsie	100 %	Chirurgie

Tableau 1.1 : Classification des anomalies et de leurs investigations [12]

1.6 La cytopathologie du sein

La cytologie (figure 1.9) est un examen complémentaire, à la mammographie et l'échographie. Elle doit être réalisée après un bilan sénologique (mammographie et échographie) pour éviter des résultats erronés. Une ponction cytologique ou cytoponction est réalisée. Elle consiste à prélever des cellules, au niveau d'une anomalie du sein à travers la peau, à l'aide d'une seringue et d'une aiguille fine d'où l'appellation ponction à l'aiguille fine. Ce prélèvement subit un examen microscopique, afin

d'identifier la nature de la lésion et de décider si un traitement est nécessaire, ou pas. Si c'est le cas, les médecins sont orientés sur le choix du traitement [14, 15, 16].

Elle est réalisée de deux manières :

- Si la tumeur est palpable, la cytologie se fera par une seringue à fine aiguille en pleine masse.
- Si elle n'est pas palpable, la cytologie se fera sous échographie guidée.

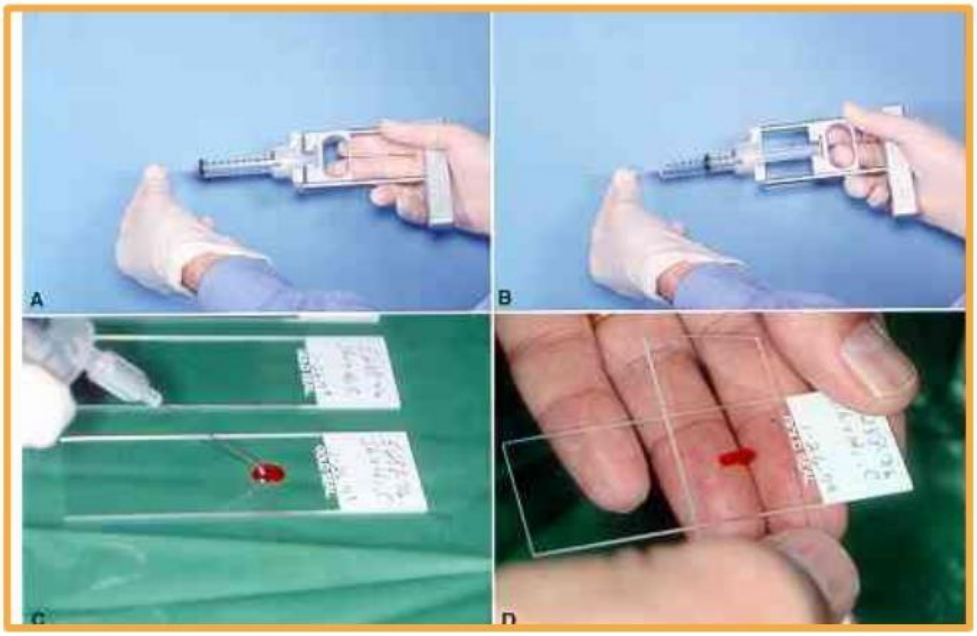


Figure 1.9 : Etalement cytopathologique [14]

1.6.1 Caractères pathologiques d'une cellule

La cellule est une unité biologique fondamentale structurelle et, fonctionnelle pour tous les êtres vivants. Elle est capable de se reproduire façon autonome [14, 16].

La cellule est composée d'une membrane plasmique, qui la délimite et qui contient le cytoplasme. Le cytoplasme est le liquide où baignent de nombreux organites. Il contient des enzymes essentiels à la survie de la cellule et des métabolites. Le matériel génétique se situe dans le noyau. C'est par le biais de la mitochondrie que l'énergie nécessaire est délivrée.

La figure 1.10, illustre la division anarchique cellulaire, qui produit un cancer.

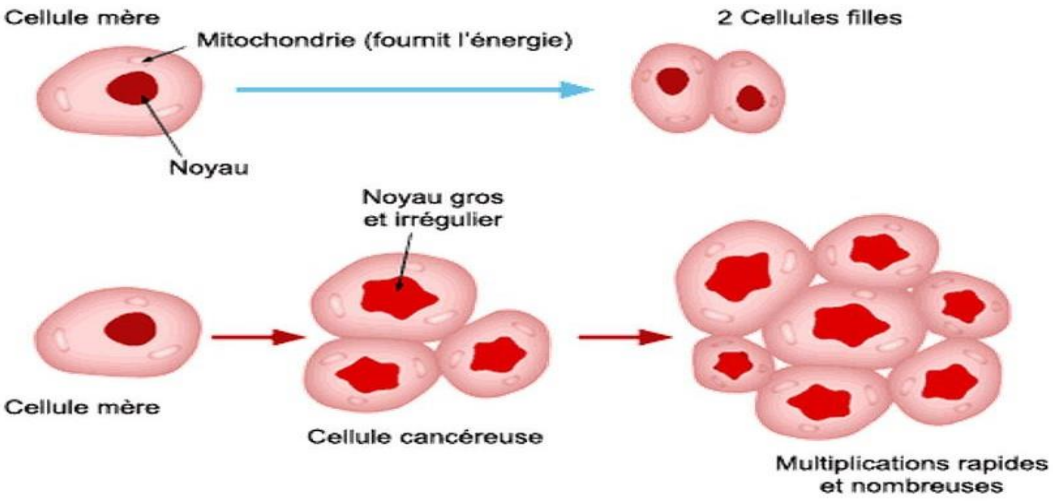


Figure 1.10 : Division cellulaire anarchique [14]

La cellule cytopathologique présente des critères de malignité (figure 1.11), touchant principalement le noyau et le cytoplasme [14, 16].

- **Anomalies du noyau**

Les anomalies de la chromatine, ou de la ploïdie, sont à l'origine de multiples anomalies nucléaires :

- Les noyaux sont de taille inégale d'une cellule à une autre ;
- Les noyaux sont foncés et denses ;
- la cellule tumorale peut avoir plusieurs noyaux ;
- La membrane nucléaire est épaissie, les contours nucléaires sont irréguliers ;
- Les nucléoles sont multiples, volumineux, irréguliers.

- **Anomalies du cytoplasme**

- Le cytoplasme est moins abondant, ce qui contribue aussi à l'augmentation du rapport nucléo-cytoplasmique.
- La taille des cellules est variable (anisocytose).
- Le cytoplasme est basophile (par augmentation de son contenu en acides nucléiques).

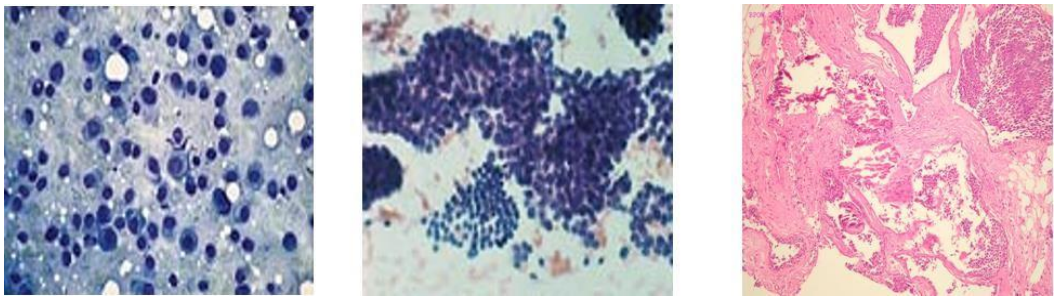


Figure 1.11 : Exemples d'images cytopathologiques [14]

1.6.2 Evaluation de la cytologie

La fiabilité du diagnostic cytologique des tumeurs mammaires, dépend fortement de l'opérateur et des catégories des lésions à analyser. Dans le cas, d'un écoulement isolé et sans tumeur palpable, la fiabilité ne dépasse pas 50 %. Lorsqu'il s'agit de tumeurs solides, la sensibilité pourrait atteindre, une valeur de 94% [16].

La cytologie est fiable à plus de 90%, mais néanmoins sa négativité n'élimine pas les faux négatifs (5% à 10% de faux négatifs).

La cytologie a donc ses limites ; mais les risques d'erreurs peuvent être minimisés si l'étude cytologique est confrontée aux données cliniques, radiologiques et aux systèmes d'aide à la décision.

1.7 Conclusion

Ce chapitre a été consacré à l'étude du contexte médical mammaire. La principale technique de dépistage a été abordée, pour comprendre l'utilité de la cytologie, dans le diagnostic du cancer du sein.

Une bonne analyse des anomalies mammaires est le but des médecins. Néanmoins, certaines lésions sont omises lors d'un examen cytopathologique, le traitement d'image devient un outil indispensable pour l'aide à la détection de cellules cancéreuses.

Chapitre 2 Apprentissage Automatique

2.1 Introduction

L'intelligence artificielle (IA) est l'une des disciplines scientifiques et d'ingénierie les plus récentes et les plus florissantes qui a modifiée durablement notre société. Ce qui était, il y a quelques années encore de l'ordre de la science-fiction, est désormais une réalité. Nous parlons avec des ordinateurs, nos téléphones nous orientent et nous indiquent le chemin le plus court (GPS), nos montres savent si nous avons suffisamment bougé dans la journée. La technique est de plus en plus intelligente, les scientifiques, ingénieurs et programmeurs deviennent des enseignants : ils « entrainent » les ordinateurs à apprendre de façon autonome.

Aujourd'hui, Le champ d'utilisation de l'apprentissage automatique est vaste. Il est intéressant pour plusieurs domaines comme les entreprises informatiques comme Google ou Microsoft, les sciences de l'ingénierie et particulièrement en imagerie médicale pour l'aide à la décision. Dans ce chapitre, nous introduisons le machine learning, ses différentes types (supervisée nous supervisée, par renforcement … etc.) et nous décrivons les principales méthodes de classification supervisée, utilisées dans le contexte de ce mémoire.

2.2 Introduction à l'apprentissage automatique

2.2.1 Principe

L'apprentissage automatique, également appelé apprentissage artificiel ou bien machine 'learning' en anglais, est un sous domaine (champ d'étude) de l'intelligence artificielle. Il consiste à donner à une machine, la capacité d'apprendre à résoudre un problème sans être humainement guidé au cours de son apprentissage et, sans devoir programmer explicitement chaque règle, pour gérer une expérience spécifique. L'idée du machine Learning est donc de résoudre des problèmes, en modélisant des comportements grâce à un apprentissage basé sur des données [17].

2.2.2 Approches de l'apprentissage automatique

Il existe différentes approches qui varient selon le type et le volume des données. Dans les paragraphes qui suivent, nous discutons des catégories de l'apprentissage automatique [17, 18].

- **Apprentissage supervisée :** on dit qu'un apprentissage est supervisé, lorsque les données d'entrée du processus, ont déjà été catégorisées. Les méthodes supervisées doivent s'en servir, pour prédire un résultat, en vue de pouvoir le faire, sur de nouveaux échantillons non catégorisés.

- **Apprentissage non supervisée** : il est beaucoup plus complexe. Le système doit détecter les similarités entre les données, qu'il reçoit et les organiser en fonction de ces dernières. L'apprentissage non supervisé est utilisé lorsque le problème nécessite une quantité massive de données non étiquetées. Par exemple, les applications des réseaux sociaux (Facebook, Twitter, Instagram), exploitent toutes des données non étiquetées.

- **Apprentissage par renforcement :** il s'agit d'un modèle d'apprentissage comportemental. L'algorithme reçoit un feedback de l'analyse des données et guide l'utilisateur, vers le meilleur résultat. Cet apprentissage diffère des autres types d'apprentissages car le système n'est pas formé avec un ensemble de données exemples. Il apprend plutôt par le biais d'une méthode d'essais et d'erreurs. Par conséquent, une séquence de décisions fructueuses aboutit au renforcement du processus, car c'est lui qui résout le plus efficacement le problème posé.

- **Apprentissage en profondeur :** apprentissage profond (deep Learning), est un type spécifique d'apprentissage automatique qui intègre des réseaux neuronaux en couches successives, pour qu'ils apprennent des données de manière itérative. Ces réseaux neuronaux complexes sont conçus pour émuler le cerveau humain, de façon que les ordinateurs puissent être entrainés pour faire face à des abstractions et des problèmes mal définis. L'apprentissage profond est utilisé généralement dans les applications de reconnaissance d'image, de communication orale et de vision numérique.

- **Apprentissage semi-supervisé**
  L'apprentissage semi-supervisé, consiste à mettre un petit ensemble d'objets étiquetés et un grand ensemble d'objets non étiquetés. Le but est d'essayer de tirer profit à la fois des données avec et sans labels, pour résoudre des tâches.

- **Apprentissage actif**
  Dans l'apprentissage actif, nous avons un petit ensemble d'objets avec pour chaque élément son label. Il faudra donc interagir avec l'utilisateur et lui demander de donner le label d'un nouvel objet, afin de mieux apprendre le modèle de prédiction.

2.3 Méthodes de classification

La classification automatique consiste à regrouper divers objets (ou individus) en sous-ensembles d'objets (les classes). Deux types de méthodes sont utilisés pour la classification, on trouve les méthodes supervisées et les méthodes non supervisées [18, 19, 20].

2.3.1 Classification non supervisée

Lorsque le type des objets dans l'image n'est pas connu, cela résulte d'un manque d'information sur la réalité du terrain. A cet effet, il existe des algorithmes de classification qui permettent de créer un regroupement de pixels similaires. Dans le cadre de la classification non-supervisé, il est généralement nécessaire d'indiquer le nombre de classe recherchées. Les textures similaires sont alors regroupées à l'intérieur d'une même classe, sans l'intervention de connaissances a priori et simplement, à partir d'estimations de similarité entre caractéristiques.

2.3.2 Classification supervisée

La classification supervisée est une technique de classification automatique où l'objectif principal, est de produire automatiquement des règles pour regrouper des objets en un certain nombre de classes, à partir d'une base de données d'apprentissage contenant des échantillons. L'intervention d'un utilisateur est indispensable, ainsi qu'une première phase d'apprentissage durant laquelle, le système apprend les caractéristiques associées à chaque classe d'une base d'apprentissage.

Il existe plusieurs types des classifieurs supervisés et non supervisés. Dans le cadre de ce mémoire, nous nous intéressons aux classifieurs supervisées, qui se basent sur une règle de décision à partir d'une image déjà catégorisée.

Dans la classification supervisée (figure 2.1), les méthodes les plus connues sont les suivantes : K-NN, SVM, et le MLP.

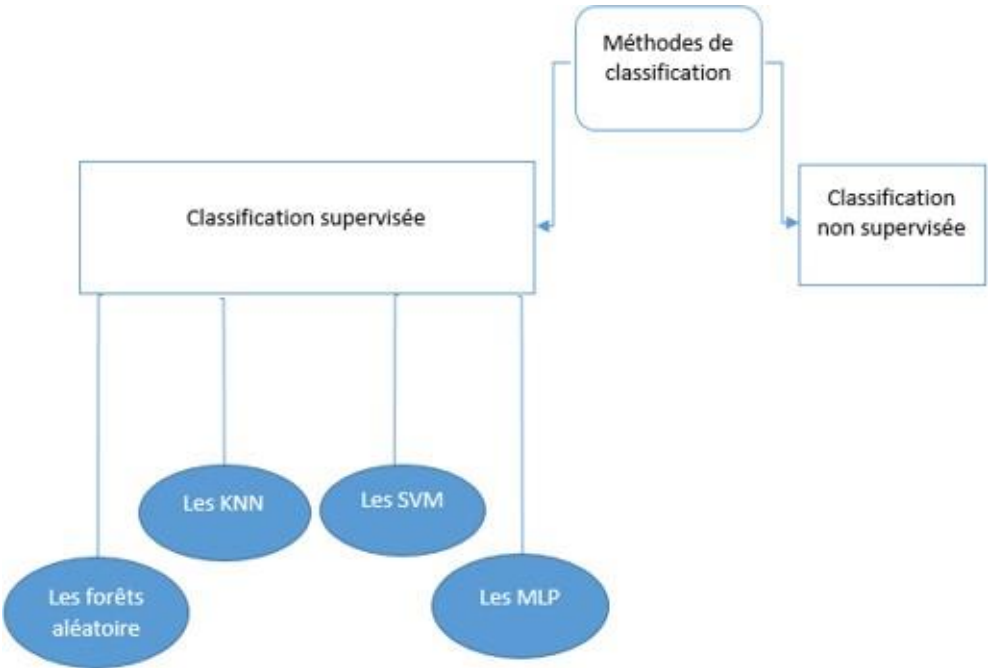


Figure 2.1 : Schéma représentatif des méthodes de classification supervisée [17]

2.4 Classifieur K-NN

L'algorithme kNN suppose que des objets similaires existent à proximité. En d'autres termes, des éléments similaires sont proches les uns des autres. L'approche des K-plus proches voisins (K-Nearest Neighbors (K-NN) en anglais), est l'un des algorithmes les plus utilisés en apprentissage automatique supervisé, pour la classification des données. Cet algorithme (figure 2.2) consiste à estimer la classe d'une nouvelle donnée d'entrée, en cherchant les k plus proches voisins qui ont été déjà classés, suivant une distance pour choisir dans ce cas, la classe des voisins majoritaires [21, 22].

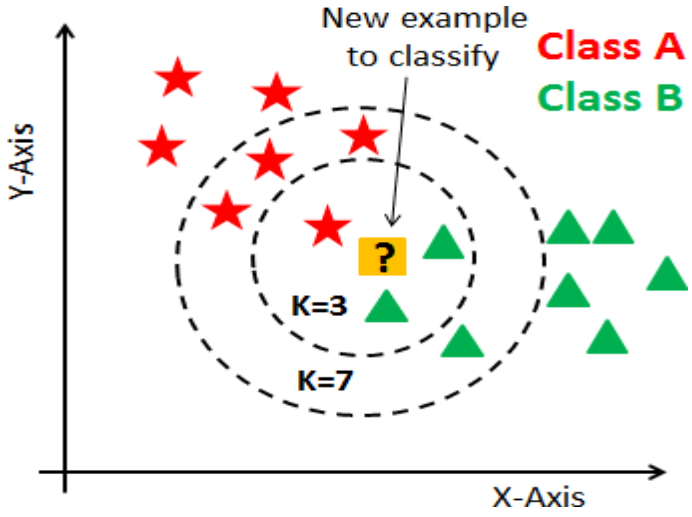


Figure 2.2 : Exemple d'illustration de k plus proches voisins [21]

Les étapes permettant de réaliser la classification par les K-NN, sont notées ci-dessous :

Début

Données en entrée

- un ensemble de données D.
- une distance d.
- Un nombre entier K.

1- Fixer le nombre de voisins k.

2- Calculer toutes les distances d'une observation X avec les autres observations du jeu de données D.

3- Détecter les k-voisins les plus proches des nouvelles données d'entrée D à classer.

4- Attribuer les classes correspondantes par vote majoritaire.

- Le choix du nombre K lors de l'implémentation de l'algorithme se fait comme suit :
  - Faire varier K.
  - Pour chaque valeur de k, il faudra calculer le taux d'erreur de l'ensemble test.
  - Retenir le paramètre k qui minimise le taux d'erreur test.

Fin

Il existe plusieurs fonctions de calculer la distance, notamment, la distance euclidienne, la distance de Manhattan, la distance de Minkowski, la distance de Hamming…etc. Généralement, on choisit la distance en fonction des types de données qu'on manipule. Ainsi pour les données du même type, la distance euclidienne est un bon candidat (équation 2.1).

La distance euclidienne entre deux points est obtenue par la formule suivante :

$$d^2(i,k) = \sum_{j=1}^{P}(x_{ij} - x_{kj})^2 \qquad 2.1$$

Où :

(i, k) : sont les deux points les plus proches.

x : Vecteur de caractéristiques de dimension p relatif à ces deux points.

➢ Avantages et inconvénients

La classification des K-NN, présente des avantages comme :

- Facilité de compréhension.
- Rapidité d'apprentissage.

Les inconvénients engendrés par ce classifieur, se résument selon ces points :

- L'estimation de ce modèle devient de mauvaise qualité quand le nombre de variables explicatives est grand.
- Pas efficace pour les jeux de données larges.

2.5 Machines à vecteurs de supports

L'approche des machines à vecteurs supports (support vector machine, SVM) est un algorithme qui a été proposé en 1995 par Vapnik [23, 24]. Il consiste à classer les données qui sont modélisées comme des points dispersées (vecteurs) dans un espace, en deux classes. Les classes sont séparées par un hyper-plan optimal, en maximisant la marge. La figure 2.3 illustre cette approche, où La distance du point le plus proche à l'hyper-plan est représentée par d.

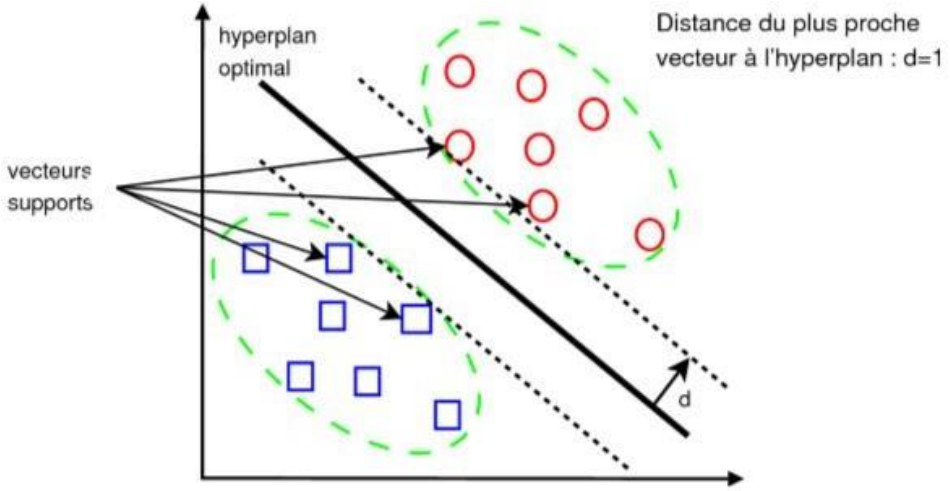


Figure 2.3 : Hyper-plan optimal et marge maximal [23]

2.5.1 Principe de fonctionnement des SVM

Le but principal de cette méthode est une classification binaire. La classification consiste à chercher un hyperplan (figure 2.4) qui sépare les exemples positifs, des exemples négatifs dans un ensemble de données ; en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. L'intérêt est donc de sélectionner les vecteurs supports qui déterminent l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont plus utiles et seuls ces vecteurs supports, sont utilisés pour classer un nouveau cas, ce qui peut être considéré comme un avantage pour cette méthode [24].

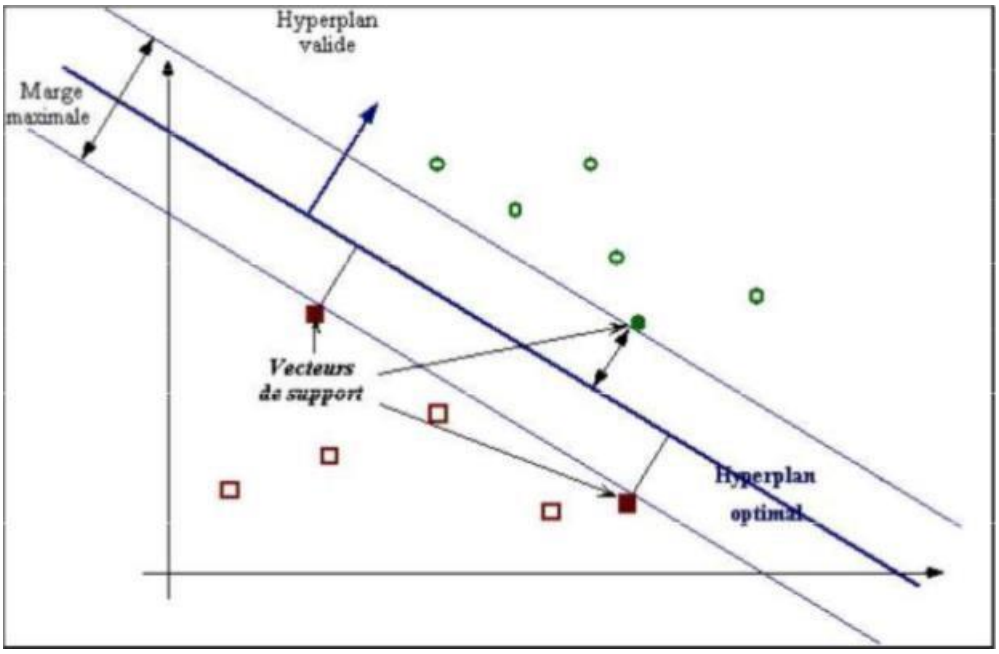


Figure 2.4 : Illustration de détermination d'un hyperplan via les SVM [23]

2.5.2 Données linéairement séparables

En apprentissage automatique, On dit qu'un classifieur est linéaire, si sa fonction de décision peut être exprimée par une fonction linéaire. Cette fonction est décrite comme suit :

$$h(x) = < w, x > + b = \sum^{n}{}_{i=1} w_i x_i \qquad 2.2$$

Où :

$h(x)$: est la fonction de décision.

w : (ϵR^n), est le vecteur des poids.

x : (ϵR^d), est un exemple ou bien une variable d'entrée.

b : (ϵR^0), est le biais ou le seuil.

n : est le nombre de composantes des vecteurs contenant les données.

d : est la dimension de l'espace.

$< , >$: Produit scalaire.

Pour savoir à quelle classe appartient un exemple (x'), il faudra prendre le signe de la fonction de décision. C'est-à-dire :

$$y = si(h(x')) \qquad 2.3$$

$si()$: est appelé le classifieur.

Géométriquement, cela consiste à chercher un hyperplan (figure 2.5) qui sépare les données des deux classes. Les points x situés sur cet hyperplan satisfont l'équation :

$$w^T + b = 0 \qquad 2.4$$

La règle de décision correspond à observer de quel côté de l'hyperplan, se trouve l'exemple. Le vecteur w définit la pente de l'hyperplan (w perpendiculaire à l'hyperplan). b translate l'hyperplan parallèlement à lui-même (figure 2.5).

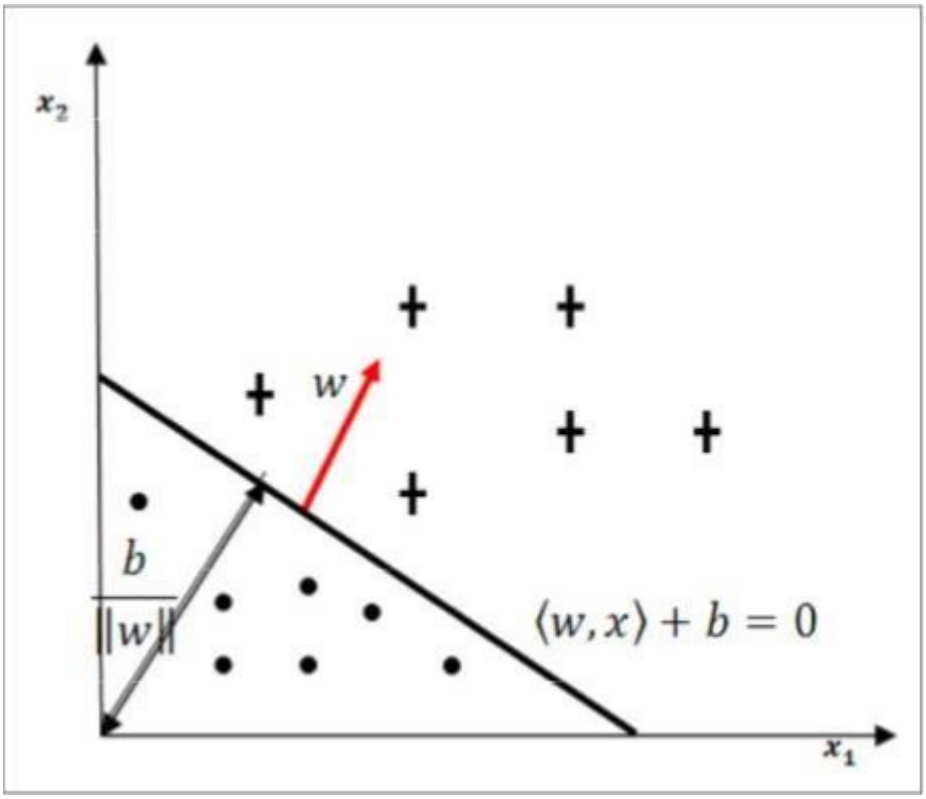


Figure 2.5 : Hyperplan séparateur [24]

- **Marge maximale de l'hyperplan**

La marge géométrique représente la distance euclidienne prise perpendiculairement entre l'hyperplan et l'exemple x_i [24]. En prenant un point quelconque x_p se trouvant sur l'hyperplan, la marge géométrique peut s'exprimer par :

$$\frac{w}{||w||} . x_i - x_p \qquad 2.5$$

L'hyperplan à marge maximale, est le modèle le plus utilisé dans les machines à vecteurs supports. L'estimation des paramètres (w^*, b^*) de l'hyperplan qui maximise la marge se fait en résolvant le problème d'optimisation suivante :

$$(w^*, b^*) = \arg m_{(w,b)}\, min_i \{ (y_i(w\, x_i + b)), ||w|| = 1\} \qquad 2.6$$

Dire que les deux classes de l'échantillon d'apprentissage sont linéairement séparables, est équivalent à dire qu'il existe des paramètres (w^*, b^*) tels que l'on a pour tout $i = 1,2, \dots, n$ [7] :

$w^* x_i + {}^* > 0 \; si \; y_i = 1$ 2.7

$w^* x_i + {}^* < 0 \; si \; y_i = -1$ 2.8

Ce qui est équivalent à :

$(w^* x_i + b > 0 \; ; \; \forall i = 1,2, \dots . ;$ 2.9

La définition consiste à dire qu'il doit exister un hyperplan, laissant d'un côté toutes les données positives et de l'autre toutes les données négatives. Dès lors, nous pouvons définir deux plans se trouvant de part et d'autre de l'hyperplan et parallèles à celui-ci, sur lesquels reposent les exemples les plus proches. La figure 2.6 illustre cette situation [8].

Dans notre définition de l'hyperplan, il est possible que différentes équations correspondant au même plan géométrique :

$(< w, x > + \; = 0)$ 2.10

a Est une constante quelconque.

Il est donc possible de redimensionner $(w^*, \; b^*)$ de telle sorte que les deux plans parallèles aient respectivement pour équations :

$w^* x_i + b^* = 1$ 2.11

$w^* x_i + b^* = -1$ 2.12

Ces deux hyperplans sont appelés hyperplans canoniques.

Ainsi la marge entre ces deux plans est égale à :

$y = \frac{2}{||w^*||}$ 2.13

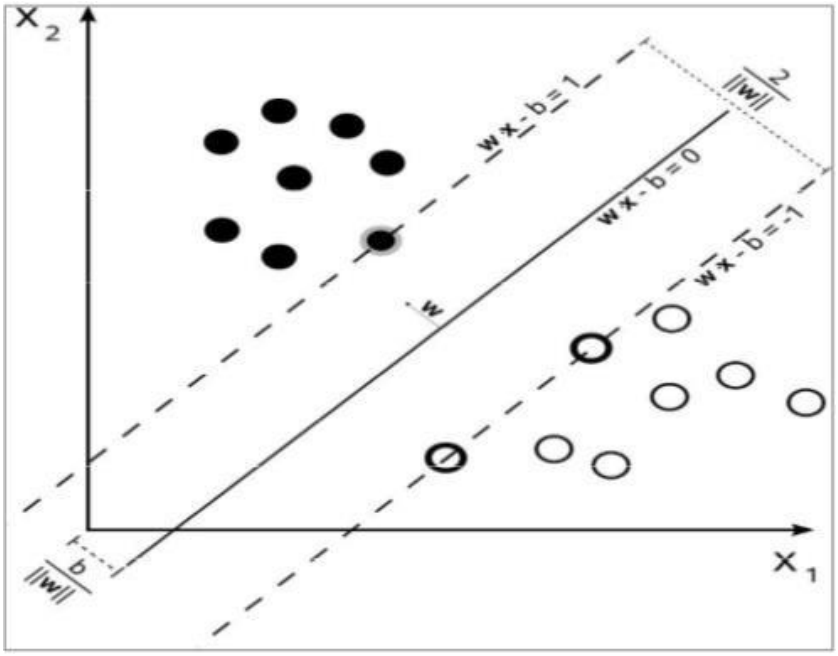


Figure 2.6 : Hyperplan canonique et marge maximale [23]

2.5.3 Données non linéairement séparables

La section vue précédemment, décrit la technique de classification par les SVM, lorsque les données sont linéairement séparables. En effet, ce n'est pas toujours le cas. Parfois, Lorsque les données des deux classes se chevauchent sévèrement dans l'espace, il est difficile de les séparer (ce jeu de données) par un simple hyperplan séparateur. Dans ce cas, l'idée est de projeter les points d'apprentissage i dans un espace de dimension plus grand, grâce à une fonction non-linéaire Ø qu'on appelle fonction noyau, choisie a priori et appliquer la même méthode d'optimisation de la marge, dans l'espace. L'espace ainsi obtenu, est appelé espace des caractéristiques (figure 2.7) ou aussi espace transformé [24].

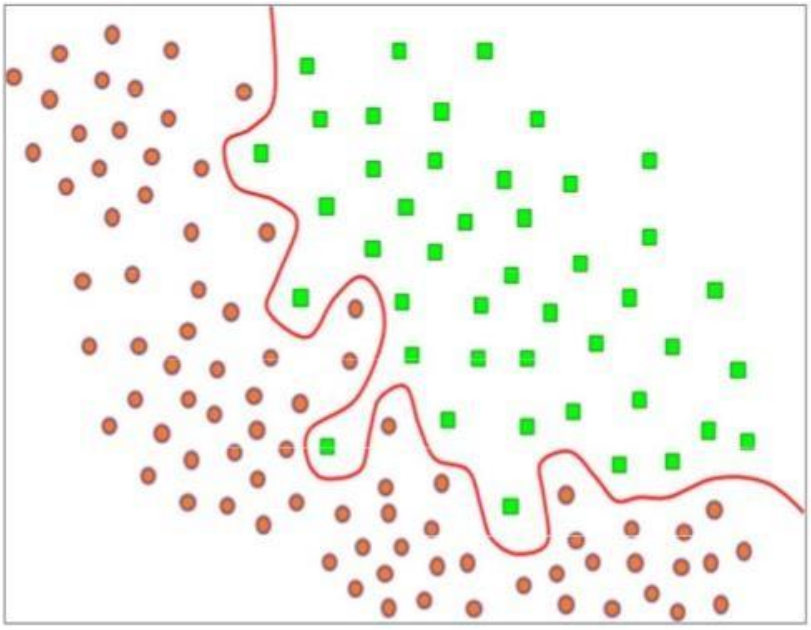


Figure 2.7 : Cas non linéairement séparable [23]

- **Marge souple**

Dans le cas de données non linéairement séparables, il est nécessaire d'utiliser les SVM de marge souple qui, en ayant une surface moins sinueuse et aussi en permettant la mauvaise classification de certaines des données d'entraînement ; améliorera la capacité de généralisation du classifieur.

La fonction de marge souple du nouveau problème de séparation optimale est reformulée comme suit :

$$min_{w,,2} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \epsilon i \ 1 \ , C \ \geq 0 \quad sous\ contraintes \qquad 2.14$$

$$, x + b \geq +1 - \epsilon i$$

$$\epsilon \geq 0\ pour\ i = 0, \ldots \ldots l$$

Où ;

ϵi : Indique à quel point l'exemple xi est loin de la marge (plus xi sera loin de la séparatrice, plus ϵi sera grand).

Le paramètre C, est un terme de pénalité introduit dans la formule par l'utilisateur. Il peut être interprété comme une tolérance, au bruit du classificateur.

- **Paramètre de pénalité**

Le paramètre de pénalité ou de régularisation sert à réduire les faux négatifs.

- **Fonction noyau**

La fonction noyau joue le rôle central de liaison des vecteurs d'entrées, à l'espace de caractéristiques de grande dimension [24].

Les choix typiques pour la fonction noyau sont :

- ***Noyau Linéaire***

$$K\ (x, x') = \langle x \cdot x' \rangle \qquad 2.15$$

- ***Noyau gaussien à base radiale (ou RBF : Radial Basis Function)***

$$(x, x_i) = \exp[\gamma\ |x - x_i|\ ] \qquad 2.16$$

- ***Noyau polynômial***

$$(x_i - x_j) = (x^T x_i + 1)^p \qquad 2.17$$

Avec

p : une constante qui spécifie le degré du polynôme.

$\gamma = \frac{1}{2\delta^2}$

δ : La variance de la fonction gaussienne.

2.5.4 Avantages et inconvénients des SVM

Les avantages des SVM se résument dans les points suivants :

- La complexité du modèle ne dépend pas de la dimension, de l'espace de caractéristiques. Les SVM sont donc particulièrement intéressants à traiter, avec des données à dimensions élevées (telles que les représentations vectorielles de texte).

- En utilisant seulement les vecteurs de support dans la frontière de décision, il est possible d'obtenir des résultats de faible densité, en traitant des grands ensembles de données.

- La complexité du classeur peut être contrôlée par le paramètre de pénalité C, en rendant la marge plus lisse ou plus sinueuse au besoin.

- Il est possible d'avoir une idée de la capacité de généralisation, du classifieur, même sans avoir vu qu'une seule donnée de test.

- Avec les SVM linéaires, C'est-à-dire, sans noyau, la frontière de décision construite Correspond à un hyper-plan dont les coefficients peuvent être interprétés comme l'importance que le classeur donne aux caractéristiques d'après leur importance pour la classification.

- L'utilisation de la SVM linéaire est particulièrement bonne pour la classification des documents de texte pour les raisons suivantes, mentionnées par la plupart des problèmes de catégorisation des textes sont linéairement séparables.

- Il y a peu de caractéristiques insignifiantes.

Parmi les inconvénients des SVM ;

- Les SVM peuvent avoir des problèmes dans le cas où beaucoup des termes sont non significatifs pour la discrimination de classe.

- Elles ne servent pas à calculer des probabilités d'appartenance à la classe.

- L'utilisation des fonctions noyaux, ne permet pas de sélectionner des caractéristiques importantes pour la classification. [10]

2.5.5 Domaines d'application des SVM

Les machines à support de vecteurs ont été appliquées avec succès dans divers domaines, tels que :

- La reconnaissance et l'identification de visages.
- La reconnaissance de caractères manuscrits et des chiffres.
- La reconnaissance vocale.
- La prédiction.
- Les SVM sont utiles dans la catégorisation de texte et d'hypertexte, car leur application peut considérablement réduire le besoin d'instances de formation marquées, dans les paramètres inductifs et traductifs standard.

- La classification des images peut également être effectuée, à l'aide de SVM [4]. Cela vaut également pour les systèmes de segmentation d'image.

- L'algorithme SVM a été largement appliqué dans les sciences biologiques et médicales [19, 24].

2.6 Réseaux de neurones artificiels

2.6.1 Principe

Les réseaux de neurones artificiels sont une technique du machine learning très populaire, qui simule le mécanisme d'apprentissage biologique, plus précisément le système nerveux humain (figure 2.8). En effet, le système nerveux humain contient des cellules appelées neurones. Ces neurones sont connectés les uns aux autres, par les axones et les dendrites. Les régions de connexion entre les axones et dendrites, sont appelés les synapses. Ce mécanisme est imité en réseaux de neurones artificiels (figure 2.9), qui contiennent des unités de calcul appelées neurones [26, 27].

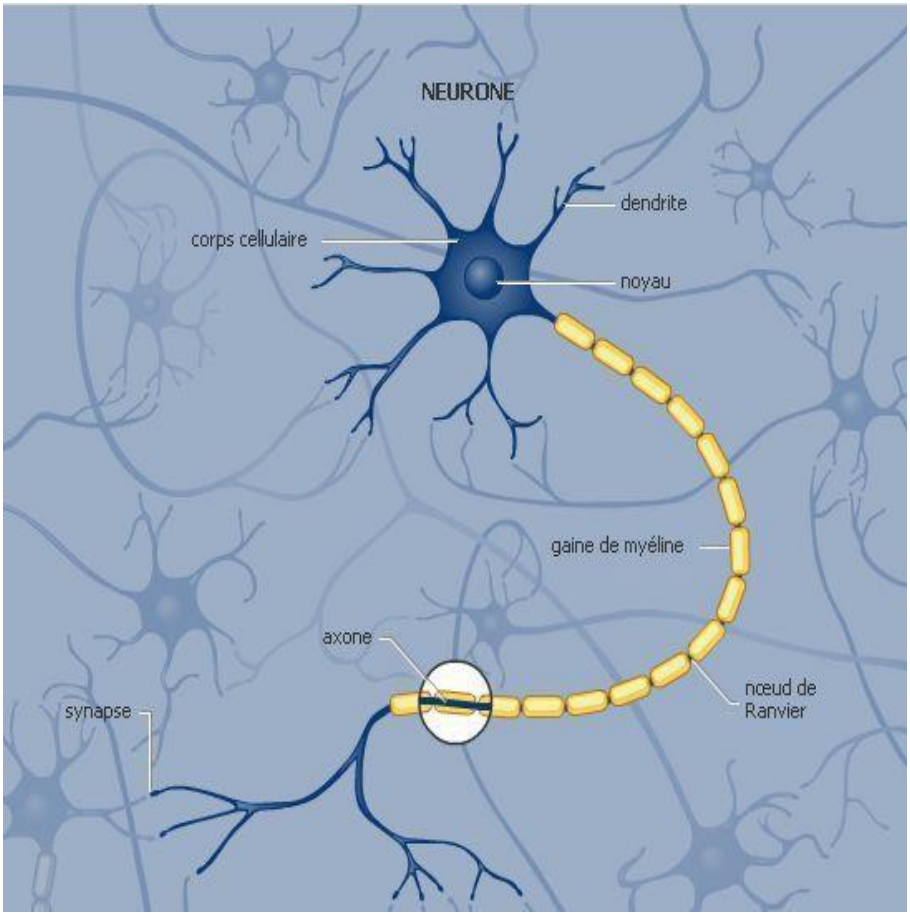


Figure 2.8 : Modèle d'un neurone biologique [28]

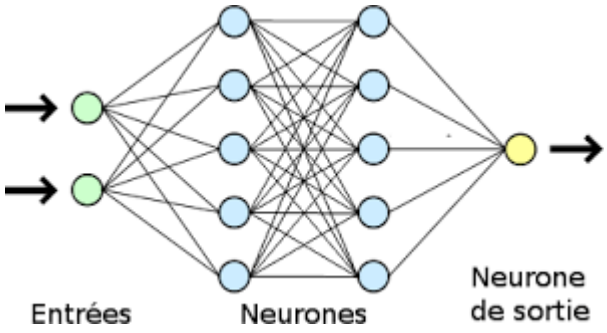


Figure 2.9 : Réseau de neurones artificiels [28]

2.6.2 Modélisation d'un neurone artificiel

Un neurone artificiel, peut être modélisé suivant le schéma de la figure 2.10.

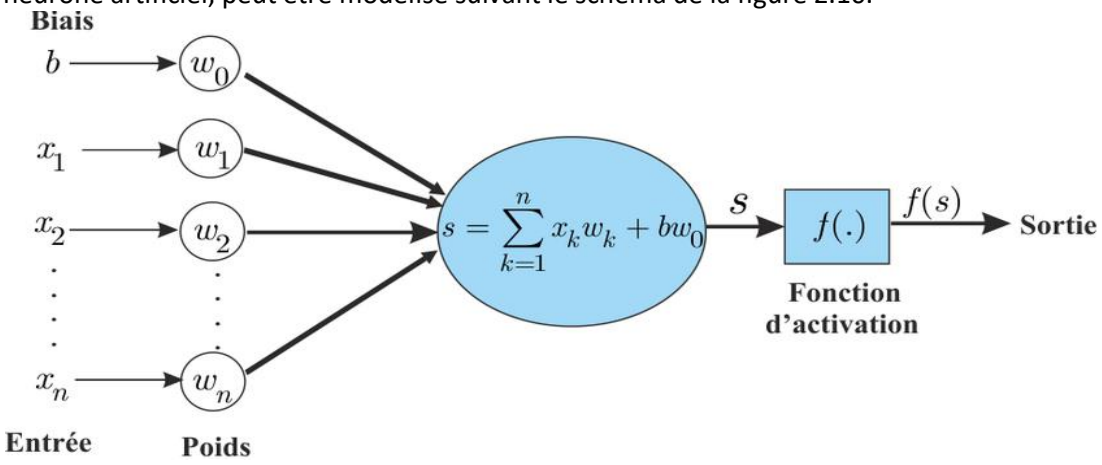


Figure 2.10 : Modélisation d'un neurone artificiel [28]

Un neurone est essentiellement constitué d'un intégrateur, qui effectue la somme pondérée de ses entrées. Le résultat de cette somme, est ensuite transformé par une fonction de transfert f qui produit la sortie du neurone [27].

Le neurone est défini comme suit :

$$s = \sum_{i=1}^{n} w_i x_i + w_0 b \qquad 2.18$$

$$Sortie = (s) \qquad 2.19$$

Chaque entrée x est connectée au neurone, via un poids synaptique w. Si w_i est positif alors l'entrée x_i est excitatrice et, inversement si w_i est négatif elle est inhibitrice.

L'entrée x_0 est en général appelé le biais (dans la figure 2.10, il s'agit de b). En effet, le biais est un neurone, pour qui la fonction d'activation est toujours égale à un [27].

2.6.3 Modèle du perceptron multicouche

Le perceptron multicouche (figure 2.11), est une amélioration du perceptron monocouche [5], comprenant une ou plusieurs couches intermédiaires dites couches cachées ; dans le sens où elles n'ont qu'une utilité intrinsèque pour le réseau de neurones et pas de contact direct avec l'extérieur. Chaque neurone n'est relié qu'aux neurones des couches précédente et suivante, mais à tous les neurones de ces couches.

Le modèle du perceptron multicouche est composé de plusieurs couches (au minimum trois couches) : une couche d'entrée (input layer), une ou plusieurs couches cachées (hidden layers) et une couche de sortie (output layer). Une couche représente un ensemble de neurones qui ne sont pas interconnectés entre eux.

Ce type de réseaux est appelé 'Feed-Forward' parce que les neurones ne sont connectés que dans un sens : les signaux provenant des entrées traversent les couches cachées pour atteindre la couche de sortie [27].

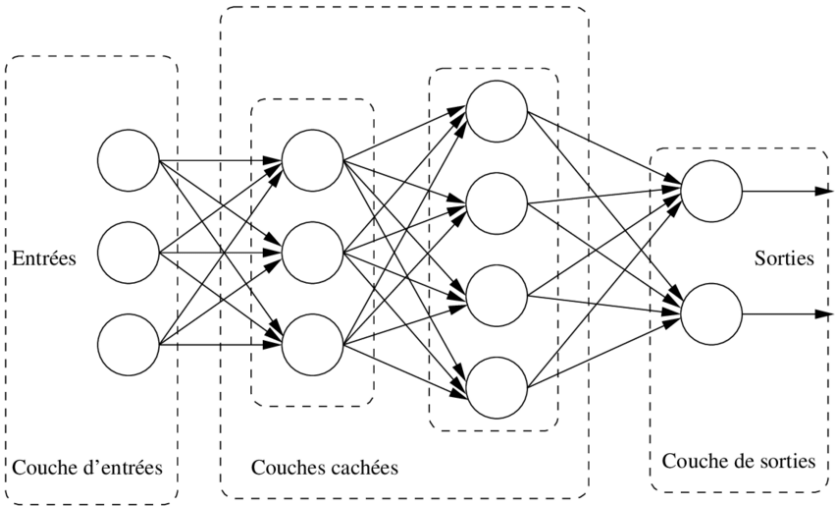


Figure 2.11 : Perceptron multicouche [28]

Le modèle du perceptron multicouche est un ensemble de neurones, organisés en couches. D'une couche à l'autre se propage le signal d'entrée jusqu'à la sortie, en activant ou non au fur et à mesure, des neurones.

Chaque couche cachée à un nombre de neurones différent comme dans la figure 2.11. La première couche cachée a 3 neurones et la deuxième couche caché a 4 neurones. Ces couches-ci peuvent elles aussi, avoir un biais ou ne pas en avoir. Son principe est d'observer l'état de la sortie par rapport, à ce

qui était attendu et de mettre à jour les liaisons entre les neurones, pour améliorer le résultat final qui sera une prédiction de la part du réseau.

Les Feed-Forward ont un comportement stable et une tolérance aux erreurs. Ils sont de loin les plus efficaces, pour aborder des problématiques concrètes donc par conséquent, les plus utilisés [28].

2.6.4 Entraînement du réseau

L'entrainement d'un réseau de neurone est effectué habituellement par l'algorithme de 'backpropagation', basé sur la descente du gradient [27].

Pour chaque échantillon (input, value), on calcule l'erreur (par exemple le carré de la différence entre la sortie du réseau de neurone, et la valeur voulue).

$L=\|value-ouput\|^2$ 2.20

Pour chaque poids w, on calcule le gradient de cette fonction L, en fonction des poids du neurone. On commence par la dernière couche, puis les couches précédentes grâce à la relation de chaine. On obtient ainsi pour chaque poids w, la valeur du gradient au point donnée par l'échantillon.

$\delta L=\partial L/\partial w(value)$ 2.21

Puis on applique l'algorithme de la descente du gradient, pour modifier légèrement les poids du réseau de neurone tels que :

$w=w+\lambda\delta L$ 2.22

Où ;

λ est appelé facteur d'apprentissage (*learning rate)*.

Plus le learning rate est faible, plus le réseau apprendra lentement. Mais un learning rate trop grand, peut empêcher la convergence des poids. En général, on choisit une valeur $\lambda \leq 0.1$

2.6.5 Fonctions d'activation

Les fonctions d'activation utilisées, sont des fonctions non linéaires [27]. Deux fonctions parmi les plus courantes sont la sigmoïde (2.23), et reLu (2.24).

$$sig(x) = \frac{1}{1+e^{-x}} \quad 2.23$$

$$re(x) = max(0, x) \quad 2.24$$

Ces différentes fonctions peuvent avoir une influence sur la performance du réseau de neurones à généraliser, à partir des données qui lui sont fournies. En effet, c'est précisément cette non linéarité, qui donne toute sa puissance au réseau de neurone et le rend capable d'approximer n'importe quelle fonction continue.

2.7 Conclusion

Dans ce chapitre ont été décrits les trois classifieurs supervisés, à adopter dans l'analyse des données cytologiques mammaires : les SVM, K-NN et le perceptron multicouche (MLP). Ceci est à cause de leurs popularités. Les K-NN sont très utiles à cause de leur facilité de compréhension et la rapidité d'apprentissage. Quant aux SVM, ils possèdent des fondements mathématiques solides ; les exemples des tests sont comparés juste avec les supports vecteurs et non pas avec tous les exemples d'apprentissage. En effet, la classification d'un nouvel exemple, consiste à voir le signe de la fonction de décision, une décision rapide**.**

Dans le prochain chapitre, nous nous intéressons à l'apprentissage profond (deep learning) qui est un sous domaine de l'apprentissage automatique.

Chapitre 3 Architectures des réseaux de neurones dans l'apprentissage profond

3.1 Introduction

Comme nous l'avons vu précédemment, l'intelligence artificielle (AI) est un domaine récent des sciences de l'ingénierie. C'est un ensemble de théories et de techniques mises en œuvre, dans le but de réaliser des machines capables de simuler l'intelligence humaine. Cependant une approche des AI serait un programme conscient, qui serait capable d'envisager toutes les possibilités et donc de prendre la meilleure décision, mais aussi d'apprendre de ses erreurs, de se perfectionner.

L'intelligence artificielle se divise en 2 parties, la première partie appelée « machine learning» est basée sur l'utilisation des statistiques, pour que la machine ait la faculté d'apprendre. La seconde partie « deep Learning » ou « apprentissage profond » se rapporte à des algorithmes capables, de s'auto-améliorer grâce à des modélisations telles que les réseaux de neurones.

Ce chapitre, est consacré principalement au deep Learning et plus précisément, aux réseaux de neurones convolutifs, les CNN.

3.2 Introduction à l'apprentissage profond

L'apprentissage profond (Deep Learning), est un sous domaine du machine Learning qui à son tour, est un sous domaine de l'intelligence artificielle. C'est un ensemble de méthodes d'apprentissage automatique s'appuyant sur les réseaux de neurones artificiels (cités précédemment) visant à imiter la profondeur des couches du cerveau, dans le sens où chaque action, est le résultat d'une longue chaîne de communications synaptiques avec de nombreuses couches de traitement [29, 30, 31].

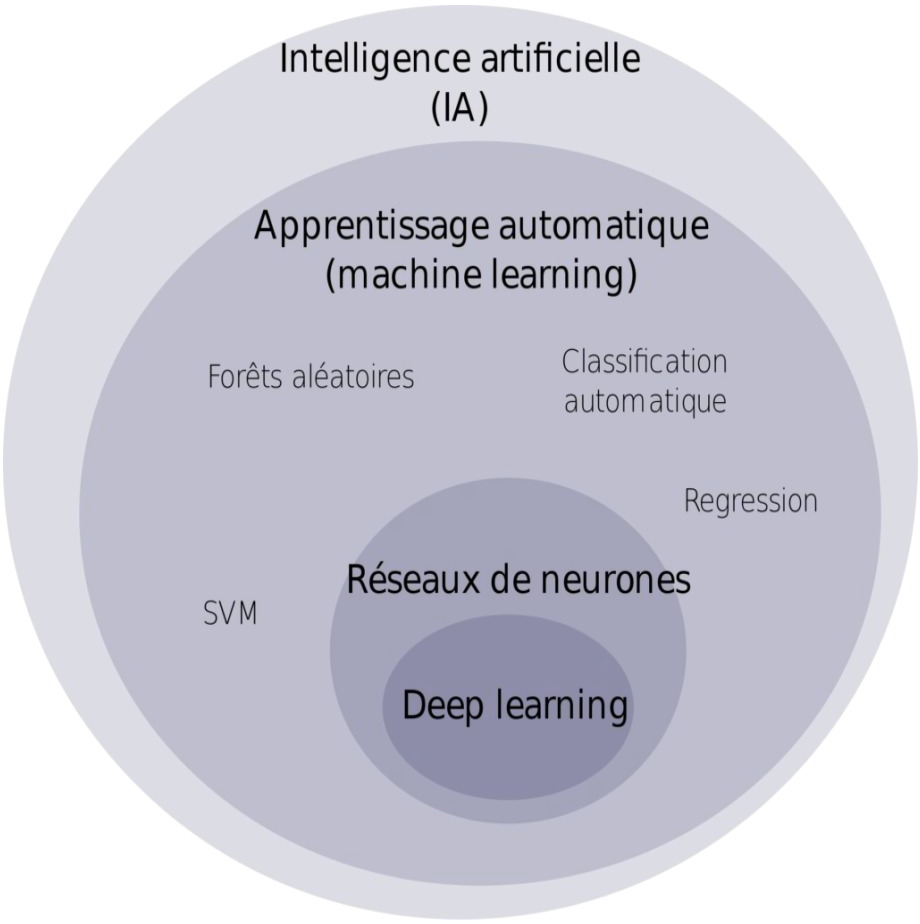


Figure 3.1 : Monde de l'intelligence artificielle [29]

L'apprentissage profond se rapporte à des algorithmes d'abstraction de haut niveau, qui permettent de modéliser les données, à partir de grands ensembles d'échantillons appris. Ces approches ont la capacité d'extraire l'information, à partir des données brutes grâce aux multiples couches de traitement et apprendre sur ces caractéristiques, petit à petit à travers chaque couche [30].

Le deep learning (le mot « deep » en anglais fait référence à la profondeur des réseaux, c'est-à-dire à leur grand nombre de couches) trouve donc son origine dans des approches relativement anciennes, mais leur exposition actuelle est due à un ensemble de nouvelles avancées terminologiques et algorithmiques, très récentes.

3.2.1 Origines du deep Learning

Avant toute chose, il est important de connaitre l'histoire de deep learning avant de découvrir son fonctionnement. Le deep learning est une terminologie nouvelle contrairement aux réseaux de neurones profonds, qu'elle désigne. Sa théorie n'est donc pas récente, même si de nouvelles méthodes algorithmiques ont permis de révéler son plein potentiel, ses fondements remontent au milieu de XXe siècle [30].

3.2.2 Réseaux du deep learning

L'apprentissage profond fonctionne suivant les modèles des réseaux de neurones, tels que :

- Le perceptron multicouche (présenté dans le chapitre 2).
- Les réseaux de neurones convolutifs (étudiés plus en détails plus tard).
- Les réseaux de neurones récurrents.

3.2.3 Domaine d'application du deep learning

Le Deep Learning est utilisé dans plusieurs domaines parmi lesquels :

- la reconnaissance d'image ;
- la traduction automatique ;
- les voitures autonomes ;
- les diagnostics médicaux ;
- la détection de fraudes…etc.

3.2.4 Fonctionnement du deep learning

Les réseaux de neurones profonds ne sont pas un simple empilement de couches de neurones. En effet, pour résoudre des problèmes complexes, il n'est pas suffisant d'ajouter toujours plus de couches. La difficulté d'apprentissage et la complexité calculatoire croissante avec le nombre de couches, sont donc parmi les problèmes les plus fréquents de réseaux de neurones.

Les solutions apporter par les recherches récentes, permettent juste de limiter ces problèmes et non pas de les régler complètement. Il est donc nécessaire de trouver des approches différentes suivant les problèmes, que l'on cherche à résoudre.

Il existe en pratique différentes types de réseaux de neurones profonds, qui visent à résoudre certains problèmes que nous mentionnons dans les sections qui suivent [31,32].

3.2.4.1 Analyse d'image

Pour analyser des images de façon pertinente, il est important d'analyser la hiérarchie des objets définis par les pixels, plutôt que chaque pixel séparément. Le type des réseaux de neurones qui analysent les images sont appelés, réseaux de neurones convolutionnels (CNN). Ces derniers, utilisent des couches de neurones particulières, appelées couches de convolutions qui analysent l'image, pour détecter ses caractéristiques importantes.

Les cas d'usages les plus fréquents en analyse d'image sont :

- La classification des images.
- La détection des objets.
- la Segmentation d'image.

Avec suffisamment de données, les réseaux de neurones actuels permettent d'obtenir des performances similaires à celle de l'humain.

Traitement de texte

Les données textuelles ont aussi à leur tour une complexité supplémentaire, liée à leurs structures ordonnées. Pour mieux comprendre une phrase dans un paragraphe, il est nécessaire de connaitre son contexte et de bien comprendre, les phrases précédentes.
Les réseaux de neurones qui permettent de traiter ce type de données temporelles, sont appelées réseaux de neurones récurrents (RNN). Ces réseaux analysent les mots du texte, un par un (voir les caractères pour certains) et gardent en mémoire les informations, qui semblent pertinentes.

En effet, les réseaux récurrents les plus fréquemment utilisés sont appelées LSTM (Long Short Terme Memory) et ils ont été inventés en 1997.

La puissance de calcul ainsi que de nouvelles méthodes, pour faciliter l'apprentissage ont encore une fois permis leurs applications à plus grande échelle depuis 2012 [32].

Les cas d'usages les plus fréquents en analyses textuelle sont :

- La classification de texte.
- L'analyse de sentiments.
- La traduction automatique.
- La modélisation du langage.

Dans certain cas d'usage comme la classification ou la traduction, les réseaux de neurones présentent des performances qui sont similaires à celles de l'humain. Néanmoins, il y a aussi des tâches pour lesquelles les réseaux, sont loin comme l'analyse de sentiments lorsque l'ironie est en jeu.

3.2.5 Deep learning pour l'imagerie médicale

Le Deep Learning suscite un réel engouement dans l'imagerie médicale. Les études abondent, tendant à montrer que l'IA pourrait demain diagnostiquer plus sûrement, une pathologie qu'un radiologue [31, 33].

Le champ d'application du deep learning en imagerie médicale, sert toutes les disciplines médicales, sans être exhaustives, nous pouvons déjà citer la radiologie, la dermatologie, l'ophtalmologie, la neurologie…etc. C'est dans ce cadre, que certains traitements en imagerie médicale, pourraient être réalisés grâce à l'apprentissage profond, tels que :

- la segmentation d'images médicales ;
- l'amélioration des systèmes de la détection précoce du cancer du sein.
- Aide au diagnostic et la classification d'image médicale.

3.3 Introduction aux réseaux de neurones récurrents

Les réseaux récurrents (ou RNN pour Recurrent Neural Networks) sont des réseaux de neurones, où l'information peut se propager dans les deux sens, y compris des couches profondes aux premières couches. En cela, ils sont plus proches du vrai fonctionnement du système nerveux humain, qui n'est pas à sens unique.

Ces réseaux possèdent des connexions récurrentes dans le sens où, elles conservent des informations en mémoire : ils peuvent prendre en compte à un instant T un certain nombre d'états passés. Pour cette raison, les RNNs sont particulièrement adaptés aux applications de traitement des séquences temporelles, comme l'apprentissage et la génération de signaux, c'est-à-dire quand les données forment une suite tout en étant dépendantes les unes des autres.

Néanmoins, pour les applications faisant intervenir de longs écarts temporels (typiquement la classification de séquences vidéo), cette mémoire à court-terme n'est pas suffisante. En effet, les RNNs « classiques » ne sont capables de mémoriser que le passé dit proche, en oubliant l'apprentissage au bout d'une cinquantaine d'itérations environ.

Ce transfert d'information à double sens rend leurs entrainements beaucoup plus compliqués et ce n'est que récemment, que des méthodes efficaces ont été mises au point comme les LSTM (long short term memory). Ces réseaux à large mémoire court-terme ont notamment révolutionné la reconnaissance de la voix, par les machines (speech recognition) ou la compréhension et la génération de texte (natural langage processing).

D'un point de vue théorique les RNNs ont un potentiel bien plus grand, que les réseaux de neurones classiques, c'est-à-dire qu'ils permettent théoriquement de simuler n'importe quel algorithme [32, 34, 35].

3.4 Introduction aux réseaux de neurones convolutifs

Les réseaux de neurones convolutifs sont une forme particulière de réseaux neuronaux multicouches, dont l'architecture des connexions est inspirée de celle du cortex visuel des mammifères [31, 35, 36].

Leurs conceptions suivent la découverte de mécanismes visuels, dans les organismes vivants. Ces réseaux de neurones artificiels sont capables de catégoriser les informations, des plus simples aux plus complexes. Ils consistent en un empilage multicouche de neurones, des fonctions mathématiques à plusieurs paramètres ajustables, qui prétraitent de petites quantités d'informations. Les réseaux convolutifs sont caractérisés par leurs premières couches convolutionnelles (généralement une à trois). Une couche convolutive, est basée comme son nom l'indique sur le principe mathématique de convolution, qui cherche à repérer la présence d'un motif (dans un signal ou dans une image par exemple).

Pour une image, la première couche convolutionnelle peut détecter les contours en objets (par exemple un cercle), la seconde couche convolutionnelle peut combiner les contours en objets (par exemple une roue) et, les couches suivantes (non nécessairement convolutionnelles) peuvent utiliser ces informations pour distinguer une voiture d'une moto. Une phase d'apprentissage sur des objets connus, permet de trouver les meilleurs paramètres en montrant par exemple à la machine des milliers d'images de chien, de voiture …etc. L'un des enjeux est de trouver des méthodes, pour ajuster ces paramètres le plus rapidement et le plus efficacement possible.

Les réseaux neuronaux convolutifs ont de nombreuses applications, dans la reconnaissance d'images, vidéos ou le traitement du langage naturel.

3.5 Avantages et inconvénients des réseaux de neurones

Les réseaux neuronaux artificiels [27] présentent plusieurs avantages et inconvénients. Nous citons quelques-uns :

3.5.1 Avantages

- Ils sont souples, c'est-à-dire qu'ils peuvent résoudre différents types de problèmes.
- Ils traitent des problèmes non structurés, sur lesquels aucune information n'est disponible à l'avance.
- Ils fonctionnent sur des données incomplètes ou bruitées. Cette lacune peut être complétée, par l'ajout de neurones à la couche cachée.

3.5.2 Inconvénients

- La difficulté de choisir les valeurs initiales des poids de connexion, ainsi que l'adaptation du pas d'apprentissage.
- La lenteur d'apprentissage.
- En cas d'erreur sur les résultats, nous n'avons aucune information sur le fonctionnement interne.

3.6 Fonctionnement des réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (convolutional neural networks (CNN)) s'inspirent du cerveau humain. Il s'agit d'un algorithme très répandu dans le deep Learning où, le modèle apprend à réaliser des tâches de classification directement à partir d'images, de textes, de sons ou de vidéos[31, 35].

Ces types de réseaux sont très utilisés pour l'indentification des modèles, dans les images pour reconnaitre des objets, des scènes, ou des visages. Ils apprennent directement à partir de données d'images et utilisent des modèles, pour la classification, en éliminant la nécessité d'effectuer une extraction manuelle des caractéristiques [6].

Les CNN utilisent un système comparable au perceptron multicouche, mais conçu pour réduire le nombre de calculs.

La structure des CNN consiste en une succession de couches : une couche d'entrée, une couche de sortie et entre les deux une couche cachée composée de nombreuses couches convolutives, couches de regroupement, couches entièrement connectées et couches de normalisation.

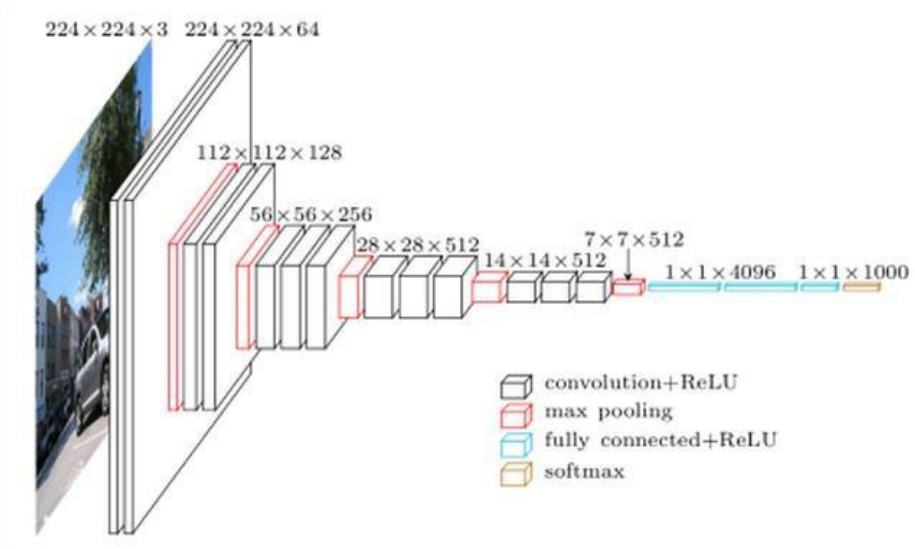


Figure 3.2 : Exemple d'un réseau de neurones convolutifs [35]

Avant de pouvoir parler des blocs de construction, nous devons d'abord comprendre ce qu'est une convolution d'image.

3.6.1 Convolution

La convolution est un outil mathématique très utilisé dans le traitement d'image. Bien que le mot « convolution » semble effrayant, celle-ci agit comme un filtrage. L'opération consiste à faire la multiplication fréquentielle, entre l'image et un noyau de taille fixe.

Le noyau va balayer toute l'image. Au début, à partir du haut de la matrice originale, le filtre se déplace de la gauche vers la droite de l'image, selon un certain pas. Lorsqu'il atteint l'extrémité, il se décalera d'un pas vers le bas et, ainsi de suite jusqu'au parcours de toute l'image (figure 3.3).

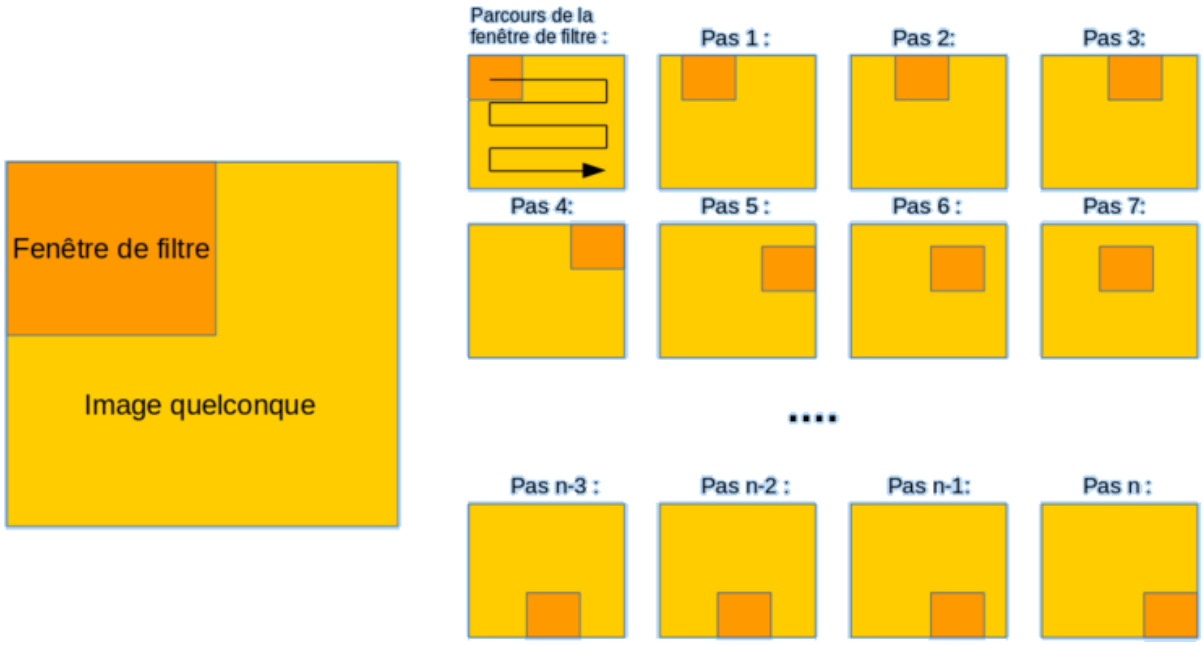


Figure 3.3 : Schéma du parcours de la fenêtre du filtre sur l'image [35]

3.6.2 Blocs de construction

Il existe quatre types de couches pour un réseau de neurones convolutifs : la couche de convolution, la couche de pooling, la couche de correction ReLu et en dernier la couche connectée [35].

3.6.2.1 Couche de convolution

La couche de convolution (figure 3.4) constitue la première couche, ainsi que la composante clé de ce type de réseaux. Elle a pour but de repérer la présence d'un ensemble de caractéristiques (features), dans les images reçues en entrée. Pour cela, un filtrage par convolution est réalisé.
Cette couche reçoit alors en entrée plusieurs images et calcule la convolution de chacune d'elles avec chaque filtre (les filtres correspondent exactement aux caractéristiques que l'on souhaite retrouver dans les images).

Pour chaque paire image-filtre, on obtient une carte d'activation (ou feature map), qui nous indique où se situent les caractéristiques de l'image. Plus la valeur est élevée, plus l'emplacement correspondant, est assigné à une caractéristique pertinente de l'image.

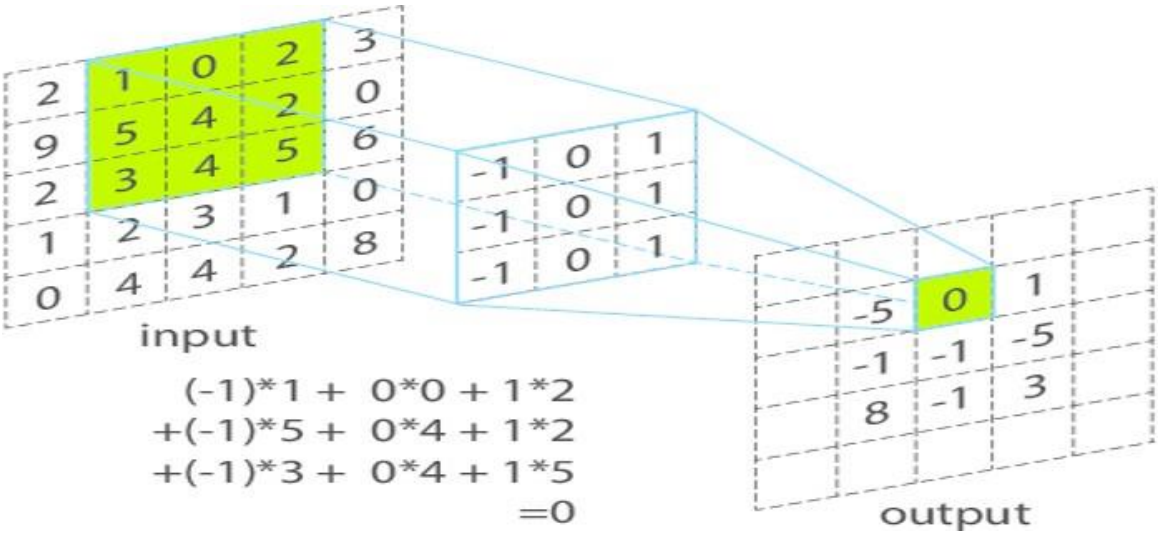


Figure 3.4 : Exemple d'une convolution d'une image avec un filtre de taille 3*3 [35]

Trois hyper paramètres permettent de dimensionner le volume de la couche de convolution (aussi appelé volume de sortie) : la 'profondeur', le 'pas' et la 'marge'.

1. 'Profondeur' de la couche : Il s'agit du nombre de noyaux de convolution (ou nombre de neurones associés à un même champ récepteur).

2. 'Le pas' contrôle le chevauchement des champs récepteurs. Plus le pas est petit, plus les champs récepteurs se chevauchent et, plus le volume de sortie sera grand.

3. 'La marge (à 0)' ou 'zero padding ' : parfois, il est commode de mettre des zéros à la frontière du volume d'entrée. La taille de ce 'zero-padding' est le troisième hyperparamètre. Cette marge permet de contrôler la dimension spatiale du volume de sortie. En particulier, il est parfois souhaitable de conserver la même surface que celle du volume d'entrée.

3.6.2.2 Couche de pooling

La couche de pooling (figure 3.5) ou couche de mise en commun, est souvent placée entre deux couches de convolution. Elle reçoit en entrée plusieurs cartes de caractéristiques (feature maps), pour appliquer à chacune d'elle, l'opération de pooling qui consiste à réduire la taille des images, tout en préservant leurs caractéristiques importantes. Il s'agit d'un sous-échantillonnage de l'image.
Dans ce cadre, l'image est découpée en cellules régulières, pour garder au sein de chaque cellule, la valeur maximale. En pratique, nous utilisons souvent des cellules carrées (ou tuiles carrées) de petite taille pour ne pas prendre trop d'informations. Les plus utilisées sont les tuiles 2*2 ou 3*3.

La couche pooling permet de réduire le nombre de paramètres et de calcul, dans le réseau, ce qui veut dire éviter le sur-apprentissage.

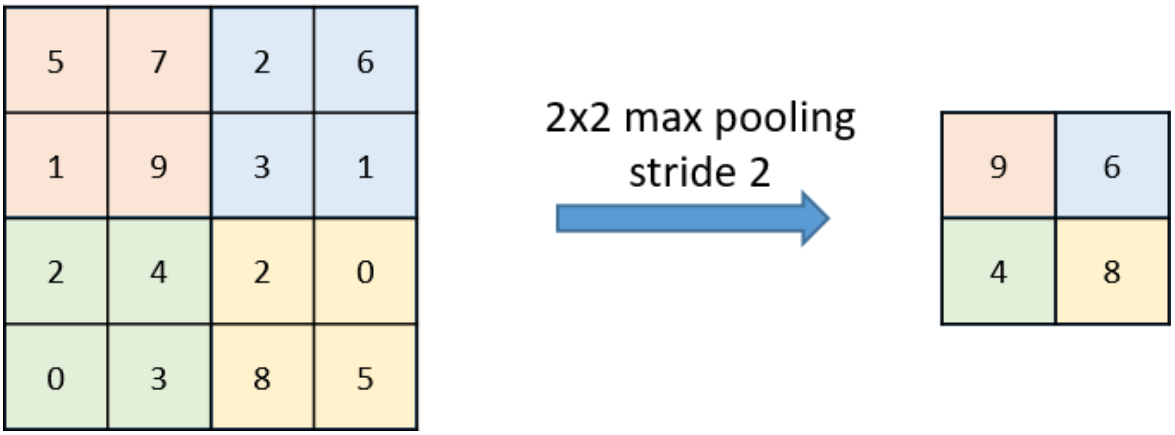


Figure 3.5 : Exemples d'un max pooling de taille 2*2 [35]

3.6.2.3 Couche ReLU

Cette couche (figure 3.6) sert à remplacer toutes les valeurs négatives reçues en entrée, par des 0. Elle joue aussi le rôle de fonction d'activation.
La fonction ReLu est une fonction mathématique réelle non-linéaire (équation 2.24).
Le résultat d'une couche ReLU est de taille identique à celle de l'entrée. La seule différence est que toutes les valeurs négatives, sont éliminées.

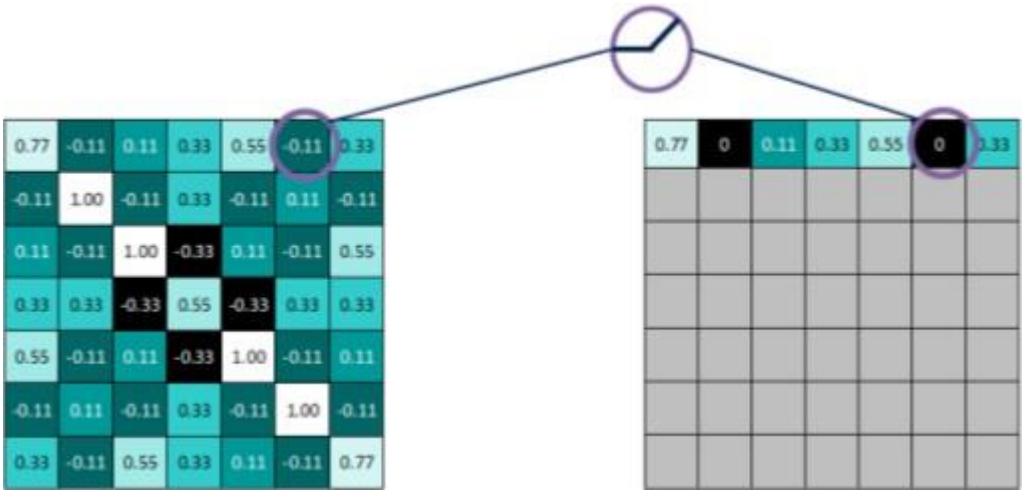


Figure 3.6 : Exemple d'application de la fonction Relu sur une image [35]

3.6.2.4 Couche fully connected

Elle constitue la dernière couche d'un réseau de neurones convolutifs. Ce type de couche reçoit un vecteur en entrée et, produit un nouveau vecteur en sortie. Pour cela, la couche applique une combinaison linéaire puis une fonction d'activation, aux valeurs reçues en entrée.

Cette dernière couche permet de classifier l'image en entrée du réseau. Elle renvoie un vecteur de taille N, où N est le nombre de classes dans notre problème de classification d'images. Chaque élément du vecteur indique la probabilité pour l'image en entrée d'appartenir à une classe.

3.6.2.5 Couche de perte (Loss)

La couche de perte spécifie comment l'entrainement du réseau pénalise l'écart entre le signal prévu et réel. Elle est normalement la dernière couche, dans le réseau. Diverses fonctions de perte adaptées à différentes tâches, peuvent y être utilisées. La perte « Soft max » est utilisée pour prédire une seule classe, parmi K classes mutuellement exclusives.

La perte par entropie croisée sigmoïde est utilisée pour prédire K valeurs de probabilité indépendante dans [0,1]. La perte euclidienne est utilisée pour régresser vers des valeurs réelles [35].

3.6.2.6 Choix des hyperparamètres

Les CNNs utilisent plus d'hyperparamètres, qu'un MLP standard. Même si les règles habituelles pour les taux d'apprentissage et des constantes de régularisation s'appliquent toujours, il faut prendre en considération les notions de nombre de filtres, leur forme ainsi que la forme du max pooling.

3.6.2.7 Choix des filtres

Comme la taille des images intermédiaires diminue avec la profondeur du traitement, les couches proches de l'entrée ont tendance à avoir moins de filtres tandis que les couches plus proches de la sortie peuvent en avoir davantage. Pour égaliser le calcul à chaque couche, le produit du nombre de caractéristiques et le nombre de pixels traités est généralement choisi pour être à peu près constant à travers les couches. Pour préserver l'information en entrée, il faudrait maintenir le nombre de sorties intermédiaires (nombre d'images intermédiaire multiplié par le nombre de positions de pixel) pour être croissante (au sens large) d'une couche à l'autre.

3.7 Modèles de quelques réseaux convolutifs

Les réseaux de neurones convolutifs, connaissent différents, s'appliquant à la classification des images [36].

3.7.1 Modèle LeNet

Il s'agit d'un réseau convolutif à 7 niveaux, qui a été appliqué à la reconnaissance des nombres manuscrits, sur les chèques numérisés. Les images d'entrée adoptées dans le cas de cette application, sont de taille de 32*32 pixels.

3.7.2 Modèle AlexNet

Le premier travail qui a popularisé les réseaux convolutifs dans la vision par ordinateur, était AlexNet (figure 3.7), développé par Alex Krizhevsky, Ilya Sutskever et Geoff Hinton. Ce CNN été soumis au défi de la base ImageNet en 2012 et a nettement surpassé ses concurrents.

Le réseau avait une architecture très similaire à LeNet, mais était plus profond, plus grand et comportait des couches convolutives empilées les unes sur les autres.

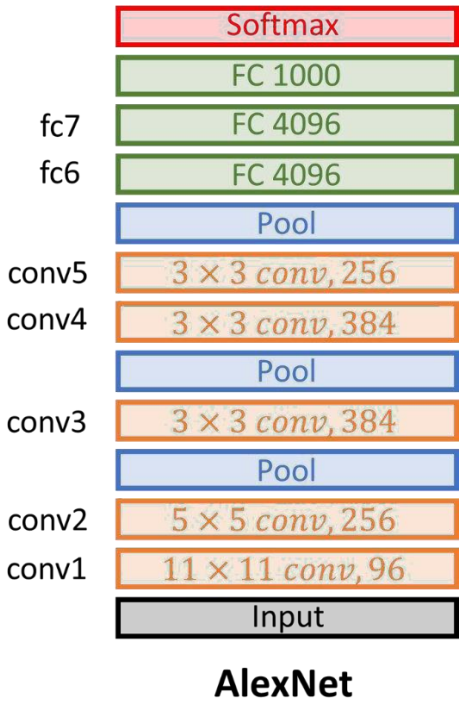


Figure 3.7: Architecture d'AlexNet [36]
(conv: couche convolutive, Pool: couche maxpooling, FC: couche fully connected)

3.7.3 VGG 16

VGG est une architecture du réseau ConvNet populaire, publié par Simonyan et Zisserman en 2014 [10]. Ce réseau montre de très bonnes performances, en se plaçant top-5 sur la classification de la base d'images ImageNet. Il se compose de 16 couches dont 13 couches de convolution et 3 fully-connected, similaire à AlexNet (figure 3.8).

3.7.4 VGG 19

Il est très similaire au VGG 16 sauf que celui-ci a 3 couches de convolutions en plus (figure 3.8).

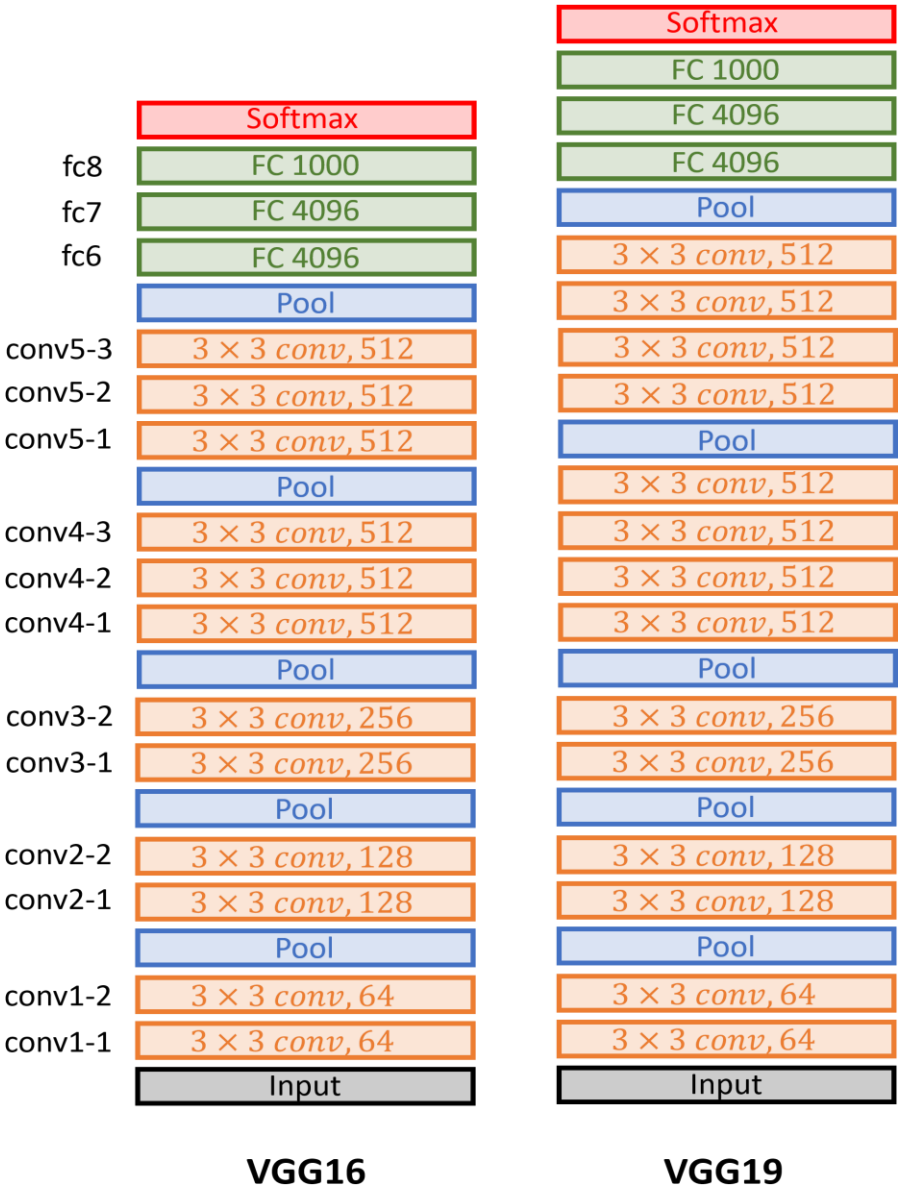


Figure 3.8 : Architectures du VGG 16 et du VGG 19 [36]

3.8 Apprentissage par transfert

Entraîner un réseau de neurones convolutif est très coûteux : plus les couches s'empilent, plus le nombre de convolutions et de paramètres à optimiser, est élevé. L'ordinateur doit être en mesure de stocker plusieurs gigaoctets de données et, de faire efficacement les calculs.

C'est pourquoi les fabricants de matériel informatique multiplient les efforts pour fournir des processeurs graphiques (GPU) performants, capables d'entraîner rapidement un réseau de neurones profond, en parallélisant les calculs.

L’apprentissage par transfert (transfer Learning) permet de faire du deep Learning, sans avoir besoin d'y passer un mois de calculs. Le principe est d'utiliser les connaissances acquises par un réseau de neurones, lors de la résolution d'un problème afin d'en résoudre un autre, plus ou moins similaire. On réalise ainsi un transfert de connaissances, d'où le nom.

En plus d'accélérer l'entraînement du réseau, le transfer Learning permet d'éviter le sur-apprentissage (overfitting). En effet, lorsque la collection d'images en entrée est petite, il est vivement déconseillé d'entraîner le réseau de neurones en partant de zéro (c'est-à-dire avec une initialisation aléatoire) : le nombre de paramètres à apprendre étant largement supérieur au nombre d'images, le risque d'overfitting est énorme !

Le Transfer Learning est une technique très utilisée en pratique et simple, à mettre en œuvre. Elle nécessite d'avoir un réseau de neurones déjà entraîné, de préférence sur un problème proche de celui qu'on veut résoudre. De nos jours, nous pouvons facilement en récupérer un sur Internet et notamment, dans les bibliothèques du Deep Learning, comme Keras. Nous pouvons exploiter le réseau de neurones pré-entraîné de plusieurs façons, en fonction de la taille du jeu de données en entrée et de sa similarité avec celui utilisé lors du pré-entraînement [36].

3.8.1 Stratégie 1 : fine-tuning total

On remplace la dernière couche fully-connected du réseau pré-entraîné par un classifieur adapté au nouveau problème et, initialisé de manière aléatoire. Toutes les couches sont ensuite entraînées sur les nouvelles images.

La stratégie 1 doit être utilisée lorsque la nouvelle collection d'images est grande : dans ce cas, on peut se permettre d'entraîner tout le réseau, sans courir le risque d'overfitting. De plus, comme les paramètres de toutes les couches (sauf de la dernière) sont initialement ceux du réseau pré-entraîné, la phase d'apprentissage sera faite plus rapidement que si l'initialisation avait été aléatoire.

3.8.2 Stratégie 2 : extraction des caractéristiques

Cette stratégie consiste à se servir des caractéristiques du réseau pré-entraîné, pour représenter les images du nouveau problème. Pour cela, on retire la dernière couche fully-connected et, on fixe tous les autres paramètres. Ce réseau tronqué va ainsi calculer la représentation de chaque image, en entrée à partir des caractéristiques, déjà apprises lors du pré-entraînement. On entraîne alors un classifieur, initialisé aléatoirement, sur ces représentations pour résoudre le nouveau problème.

Cette stratégie 2 doit être utilisée, lorsque la nouvelle collection d'images est petite et similaire aux images de pré-entraînement. En effet, entraîner le réseau sur aussi peu d'images est dangereux puisque le risque d'overfitting est important.

3.8.3 Stratégie 3 : fine-tuning partiel

Il s'agit d'un mélange des deux premières. On remplace à nouveau la dernière couche fully-connected par le nouveau classifieur initialisé aléatoirement et, on fixe les paramètres de certaines couches du réseau pré-entraîné. Ainsi, en plus du classifieur, on entraîne sur les nouvelles images les couches non-fixées, qui correspondent en général aux plus hautes du réseau.

On utilise cette stratégie lorsque la nouvelle collection d'images est petite, mais très différente des images du pré-entraînement.

3.9 Avantages des CNN

Les CNN, utilisés principalement en imagerie [35, 36], présentent de nombreux avantages que nous citons ci-dessous :

- Ils détectent automatiquement les fonctionnalités importantes, sans aucune supervision humaine.
- Ils minimisent les calculs par rapport aux réseaux neuronaux réguliers.
- La convolution simplifie le calcul dans une large mesure, sans perte de données.
- Ils sont parfaits pour la gestion de la classification d'images.

3.10 Quelques travaux sur la détection des pathologies mammaires

Les approches menées pour la détection des pathologies mammaires, moyennant le deep Learning sont multiples et variées. Nous citons brièvement certains travaux. Un système d'aide au diagnostic assisté par ordinateur, entièrement intégré pour les mammographies numériques à rayons X via la détection, la segmentation et la classification par apprentissage profond, a été réalisé par Mugahed et al [37], en 2018. Dans ce travail, pour détecter la masse mammaire, à partir de mammographies entières, les auteurs ont utilisé la méthode You-Only-Look-Once (YOLO), pour la segmentation de masses, un réseau convolutif pleine résolution (FrCN). Un réseau neuronal convolutif profond (CNN) est utilisé pour reconnaître la masse et la classer, comme bénigne ou maligne. Une multiclassification du cancer du sein à partir d'images histopathologies avec le deep Learning, est réalisée par Zhongyi et al [38], en 2017. Dans ce travail il a été question de la multiclassification du cancer du sein, ce qui signifie l'identification des classes subordonnées du cancer (carcinome canalaire, fibroadénome,

carcinome lobulaire…etc.) grâce à un modèle d'apprentissage profond structuré. Sharma et Kumar [39], ont analysé les données cytologiques du cancer du sein, via un modèle des CNN. La précision obtenue, dépasse largement les 95%.

3.11 Conclusion

Dans ce chapitre, nous avons présenté les notions de base du deep Learning, ainsi que quelques-unes de ses architectures. Nous avons concentré notre chapitre, davantage sur les CNN, puisqu'ils sont adoptés dans le cas de notre étude.

Chapitre 4 Mise en œuvre et résultats de BreastCytoLearn

4.1 Introduction

L'objectif de notre étude est de proposer une méthode de classification des images cytologiques mammaires, dans le but de catégoriser les cas malins et bénins, suivant les approches de l'apprentissage automatique et profond.

4.2 Environnement de travail

L'environnement de développement de notre système 'BreastCytoLearn', a été réalisé en utilisant le langage python, sous environnement anaconda.

4.2.1 Matériel utilisé

Notre travail a été réalisé avec un ordinateur portable de marque Acer, doté d'un processeur i3-3217U CPU@ 1.80 GHz et d'une mémoire vive RAM d'une capacité de 4 GB.

4.2.2 Langage de programmation

Le système 'BreastCytoLearn' a été développé avec le langage de programmation Python version 3.7.7. En effet, Python est un langage de programmation de haut niveau (open source) qui a été créé par le programmeur Guido van Rossum en 1991 [40]. C'est un langage de programmation interprété, qui ne nécessite donc pas d'être compilé pour fonctionner.

Il existe deux types de langages de programmation :

- **Interprété** : son utilisation nécessite un interpréteur. Il a l'avantage d'être très immédiat et l'inconvénient, d'être plus lent.

  Interpréteur

  Source → Résultat
- **Compilé** : son implémentation exécutable requiert un compilateur. Il a l'avantage d'être rapide, mais nécessite une première étape de compilation.

  Compilateur exécuteur

  Code source → code machine → résultat

Un programme ” interpréteur ” permet d’exécuter le code Python, sur n’importe quel ordinateur. Ceci permet de voir rapidement les résultats d’un changement, dans le code. En revanche, ceci rend ce langage plus lent, qu’un langage compilé comme le C.

Python permet aux programmeurs de se focaliser sur ce qu’ils font plutôt, que sur la façon dont ils le font. Ainsi, écrire des programmes prend moins de temps que dans un autre langage. Il s’agit d’un langage idéal pour les débutants.

4.2.3 Principaux avantages du langage python

Le langage Python doit sa popularité à plusieurs avantages, qui profitent aussi bien aux débutants qu’aux experts [40].

- Il est facile à apprendre et à utiliser, car ces caractéristiques sont peu nombreuses. Ce qui permet de créer des programmes rapidement et, avec peu d’efforts.
- Il contient une syntaxe qui est conçue pour être lisible et directe.
- C’est un langage qui fonctionne sur tous les principaux systèmes d’exploitation et plateformes informatiques.
- Il est utilisable pour l’automatisation ainsi que la création des logiciels de qualité professionnelle.

4.2.4 Différences entre python 2 et python 3

On distingue deux versions de Python : Python 2 et Python 3. La différence entre ces deux versions est multiple [40].

Python 2.x est l’ancienne version, qui continuera d’être supportée et donc de recevoir des mises à jour officielles, jusqu’en 2020. Après cette date, elle continuera d’ailleurs sans doute de subsister, de façon non officielle.

Python 3.x est la version actuelle du langage.

Elle apporte de nombreuses fonctionnalités nouvelles et très utiles, telles qu’un meilleur contrôle de concurrence et un interpréteur plus efficace. L’adoption de Python 3 a été longtemps ralentie, par le manque de bibliothèques tierces prises en charge. Un grand nombre d’entre elles, n’étaient compatibles qu’avec Python 2, ce qui rendait la transition compliquée.

Cependant, ce problème est aujourd’hui pratiquement résolu et il reste peu de raisons, valables de continuer à utiliser Python 2.

4.2.5 Python pour le Deep Learning et le Machine Learning

Le principal cas d'usage du Python est le scripting et l'automatisation. En effet, ce langage permet d'automatiser les interactions, avec les navigateurs web ou les GUI d'applications [40, 41].

Cependant, le scripting et l'automatisation sont loin d'être les seules utilités, de ce langage. Il est aussi utilisé pour la programmation d'applications, pour la création de services web ou de REST API, ou encore pour la métaprogrammation et pour la génération de codes.

Par ailleurs, ce langage est aussi utilisé dans de sciences des données et le domaine du Machine Learning. Avec l'essor de l'analyse de données, dans toutes les industries, c'est d'ailleurs devenu l'un de ses principaux cas d'usage.

La grande majorité des bibliothèques utilisées pour le Machine Learning ont des interfaces Python. Ainsi, ce langage est devenu l'interface de commande de haut niveau, la plus populaire pour les bibliothèques de Machine Learning et autres algorithmes numériques. De nombreux ouvrages d'initiation sont disponibles sur le Web.

Enfin, les entreprises spécialisées dans la robotique comme Aldebaran [31,33] se servent de ce langage pour programmer leurs robots. L'entreprise rachetée par Softbank [31,33] a choisi ce langage de programmation, afin de faciliter la conception d'applications par des entreprises tierces et des amateurs.

4.3 Environnement de python

Le langage python peut fonctionner sous différents environnements. Dans le cas de cette étude c'est sous Anaconda et Spyder, que nos approches ont été développées [41, 42, 43].

4.3.1 Anaconda

Anaconda est une distribution libre et open source des langages de programmation python et R appliquée au développement d'applications dédiées aux sciences des données et a l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique…). Cet environnement vise à simplifier la gestion des paquets et de déploiement. En effet, Anaconda comprend plus de 250 paquets populaires en sciences des données, adaptés pour Windows, Linux et macOS. Les versions de paquetages sont gérées par le système de gestion de paquets conda [42].

4.3.2 Spyder

Spyder (nommé Pydee dans ses premières versions) est un environnement de développement intégré (IDE), open source crée par Pierre Raybaut en 2008 [43]. Il est multiplateforme et est sous licence, non copyleft (un logiciel libre non copyleft est diffusé par son auteur, avec la permission de le redistribuer et de le modifier). Depuis 2012, c'est la communauté Python scientifique qui maintient Spyder grâce à leur contribution. Il possède les fonctionnalités de base (Coloration syntaxique, auto complétion...) et intègre de nombreuses bibliothèques d'usage scientifique telle que Matplotlib, NumPy, SciPy, IPython… etc...

4.3.3 Bibliothèques du machine learning

Si Python s'est érigé comme le meilleur langage de programmation, c'est grâce à ses différents packages et bibliothèques de science des données.

Nous présentons dans les sections qui suivent, les bibliothèques les plus populaires et utilisées en python pour le Machine Learning et le Deep Learning.

4.3.4 NumPy

NumPy [44] est une bibliothèque python très populaire, pour le traitement de grands tableaux multidimensionnels et de matrices. A l'aide d'une grande collection de fonctions mathématiques de haut niveau, cette bibliothèque facilite le calcul numérique.

L'objet principal de NumPy est le tableau homogène multidimensionnel. C'est une table d'éléments ou de nombres du même type de données, indexés par un ensemble d'entiers positifs. Les dimensions sont appelées des axes et les nombres d'axes, des rangs. Son utilisation permet d'améliorer considérablement les performances et accélérer le temps d'exécution.

- ***Que peut-on faire avec NumPy ?***
  - Opérations de tableaux de base : Ajouter, multiplier, couper, trier, indexer.
  - Opérations de tableaux avancés : diviser en section.
  - Travailler avec l'algèbre linéaire.

4.3.5 Pandas

Pandas signifie Python Data Analysis Library [45]. C'est une structure de données tabulaire bidimensionnelle à taille variable, potentiellement hétérogène avec des axes étiquetés (lignes et colonnes). Elle se compose de trois composants principaux, les données, les lignes et les colonnes.

Pandas est conçu pour la manipulation rapide et facile de données : la lecture, l'agrégation et la visualisation.

Pandas prend en entrée les données d'un fichier CSV (le cas de nos données), ou TSV ou encore une base de données SQL. A savoir que la base de données est très similaire à un tableau dans un logiciel statistique, comme Excel.

- ***Que peut-on faire avec Pandas ?***
    - Indexation, manipulation, renommer ou, tri des données.
    - Mise à jour, ajout, suppression des colonnes d'une base de données.
    - Détection de fichiers manquants et, gestion des données manquantes.
    - Représentation graphique des données.

4.3.6 SciPy

La bibliothèque SciPy [46] est l'un des paquets de base qui composent la pile SciPy. En effet, il existe une différence entre SciPy Stack et la bibliothèque SciPy. SciPy s'appuie sur l'objet NumPy et fait partie de la pile, qui comprend des outils comme Matplotlib, avec des outils supplémentaires.

Cette bibliothèque contient des modules pour « des routines » mathématiques efficaces comme l'algèbre linéaire, l'interpolation, l'optimisation, l'intégration et les statistiques. La fonctionnalité principale de cette librairie est construite sur NumPy et ses tableaux.

- ***Quand l'utiliser ?***

SciPy utilise des tableaux comme structure de données de base. Il dispose de divers modules, pour effectuer des tâches communes de programmation scientifique, comme l'algèbre linéaire, l'intégration, le calcul, les équations différentielles ordinaires et le traitement du signal.

4.3.7 Matplotlib

Matplotlib [46] est une bibliothèque de tracés 2D, utilisée pour créer des graphiques. Un module pyplot facilite le traçage, car il fournit des fonctionnalités pour contrôler les styles de ligne, les propriétés de polices, les axes …etc.

- ***Que peut-on faire avec Matplotlib ?***

Matplotlib peut représenter les données sous plusieurs visualisations :

- Graphiques linéaires.
- Graphiques de dispersion.
- Graphiques de surface.
- Graphiques à barres et histogrammes.
- Graphiques à secteurs (pie charts).
- Spectrogrammes.

- Graphiques à contours.

4.3.8 Scikit Learn

Scikit Learn [47] est une robuste bibliothèque d'apprentissage automatique, pour python. En effet, elle dispose d'algorithmes du machine learning. Elle est construite sur la base de deux bibliothèques, à savoir NumPy et SciPy. Cette librairie prend en charge la plupart des algorithmes d'apprentissage supervisé et non supervisée. Elle peut aussi être utilisée pour l'exploration et l'analyse de données, ce qui en fait un excellent outil pour débuter avec le ML.

- ***Que peut-on faire avec Scikit Learn ?***
  - Classification : reconnaissance d'image.
  - Regroupement.
  - Régression : regroupement des résultats de l'expérience.
  - Réduction dimensionnelle : visualisation, efficacité accrue.
  - Choix du modèle : amélioration de la précision grâce au réglage des paramètres.
  - Prétraitement : préparer les données d'entrée, comme un texte pour le traitement avec des algorithmes d'apprentissage machine.

4.3.9 TensorFlow

TensorFlow [48, 49] est une bibliothèque d'intelligence artificielle, qui aide les développeurs à créer des réseaux neuronaux à grande échelle avec de nombreuses couches, en utilisant des graphiques de flux de données. TensorFlow facilite également la création de modèles du deep learning et permet un déploiement facile, des applications alimentées en ML.

Il est très efficace lorsqu'il s'agit de la classification, la perception, la compréhension, la découverte, la prévision, et la création de données.

- ***Que peut-on faire avec TensorFlow ?***
  - Reconnaissance vocale.
  - Analyse de sentiments.
  - Applications textuelles.
  - Reconnaissance faciale.
  - Détection vidéo…etc.

4.3.10 Keras

Keras [50] est l'Api de haut niveau de TensorFlow, pour le développement et la formation des codes des réseaux de neurones profonds. C'est une bibliothèque neuronale open-source en Python. Avec Keras la modélisation statistique et le travail avec les images ou, le texte est beaucoup plus facile. Le codage est simplifié pour l'apprentissage profond.

La différence entre Keras et TensorFlow est que Keras, est une bibliothèque consacrée aux réseaux neuronaux python et que TensorFlow, est une bibliothèque open-source réalisant diverses tâches d'apprentissage machine.

- ***Que peut-on faire avec Keras ?***
  - Détermination du pourcentage de la précision.
  - Détermination de la fonction de perte.
  - Création des couches de fonctions personnalisées.
  - Traitement intégré des données et des images.

4.3.11 Seaborn

Seaborn [51] est une bibliothèque de visualisation de données, en python basée sur matplotlib. Elle fournit une interface de haut niveau, pour dessiner des graphiques statistiques informatifs.

- ***Que peut- on faire avec seaborn ?***

- Prise en charge spécialisée de l'utilisation de variables catégorielles, pour afficher des observations ou des statistiques agrégées.
- Visualisation des distributions univariées ou bivariées, pour les comparer entre des sous-ensembles de données.
- Vues pratiques sur les structures globales, des ensembles de données complexes.
- Structuration des grilles multi tracés, qui nous permettent de créer facilement des visualisations complexes.
- Contrôle concis sur le style des figures matplotlib, avec plusieurs thèmes intégrés.
- Outils pour choisir des palettes de couleurs qui révèlent fidèlement les motifs des données.

4.4 Système de 'BreastCytoLearn'

Le système 'BreastCytoLearn', est réparti en deux parties (figure 4.1). La première, se focalise sur la classification des données cytologiques mammaires, suivant des classifieurs traditionnels, générés par le perceptron multicouche, les K plus proches voisins, ainsi que les SVM. La seconde partie, est consacrée à la réalisation d'une approche par apprentissage profond, basée sur les CNN.

L'analyse des données cytologiques par 'BreastCytoLearn' vise à observer les caractéristiques les plus utiles, pour prédire le cancer malin ou bénin.

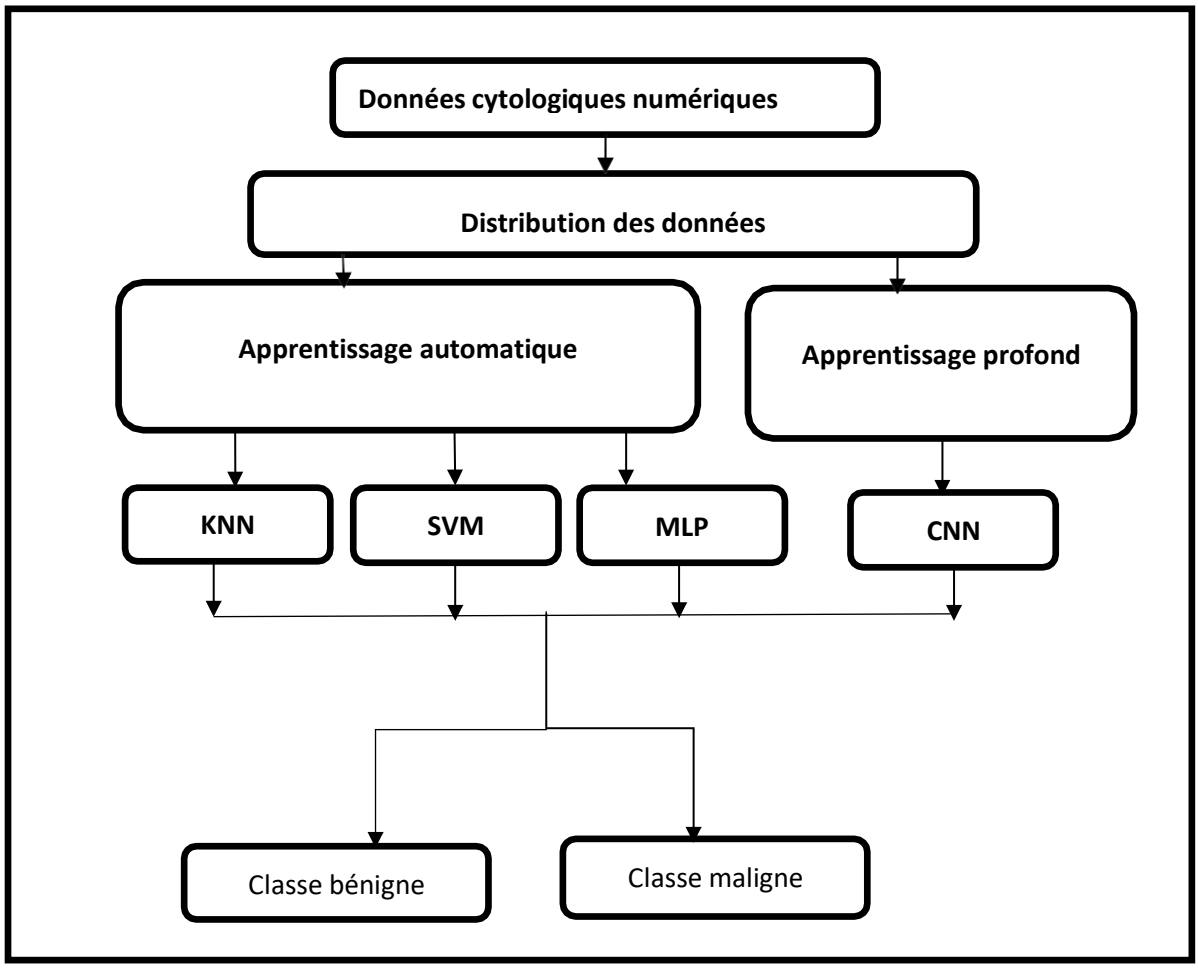


Figure 4.1 : Architecture de 'BreastCytoLearn'

4.4.1 Base de données cytologiques de Wisconsin

La base de données Wisconsin est un ensemble de données cytologiques du cancer du sein du Wisconsin, créée par William H. Wolberg, médecin à l'hôpital de l'Université du Wisconsin à Madison, Wisconsin, aux États-Unis [52].

Pour créer l'ensemble de données, le Dr Wolberg a utilisé des échantillons cytologiques, prélevés sur des patientes ayant des masses mammaires. Dix paramètres morphologiques et texturales ont été extraites, à partir des noyaux cellulaires, pour caractériser les images cytologiques. L'analyse utilise un ajustement des paramètres déterminés, pour calculer la valeur moyenne, la valeur extrême et l'erreur standard de chaque caractéristique pour l'image, renvoyant un vecteur de 30 valeurs réelles.

Le fichier des données sous excel (figure 4.2), présente des informations sur chacune des images de la base, qui comprend 357 classes bénignes et 257 classes malignes (figure 4.3).

- Informations d'attributs
- Numéro d'identification

- Diagnostic (M = malin, B = bénin)

Dix caractéristiques à valeur réelle sont calculées, pour chaque noyau cellulaire :

- rayon (moyenne des distances du centre aux points du périmètre)
- texture (écart type des niveaux de gris)
- périmètre
- Aire
- Uniformité (variation locale des longueurs de rayon)
- compacité (périmètre ^ 2 / surface - 1,0)
- concavité (sévérité des parties concaves du contour)
- points concaves (nombre de parties concaves du contour)
- symétrie
- dimension fractale.

La moyenne, l'erreur standard et les valeurs les plus élevées (moyenne des trois valeurs les plus élevées) de ces caractéristiques ont été calculées pour chaque image, ce qui donne 30 caractéristiques.

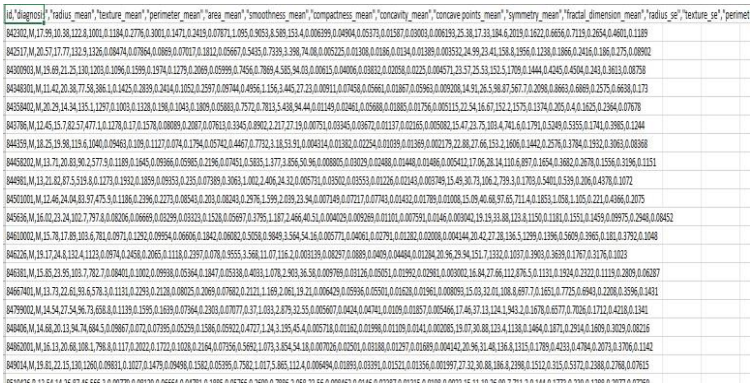


Figure 4.2 : Exemplaire du fichier original des données cytologiques [19]

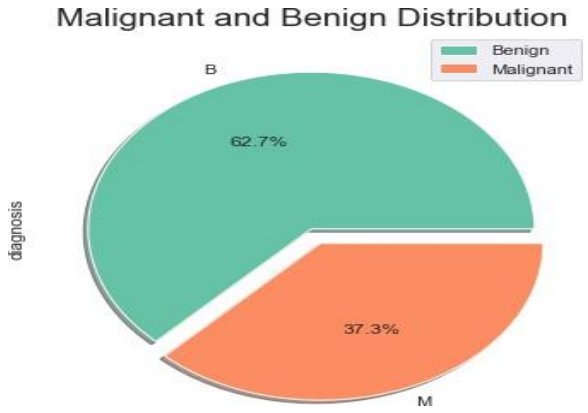


Figure 4.3 : Distribution des classes cytologiques du Wisconsin

4.4.2 Préparation des données

La préparation des données est une étape essentielle pour le problème d'analyse de données. Il faut donc préparer les données de manière à exposer au mieux la structure du problème, aux algorithmes d'apprentissage automatique que nous allons utiliser. Car certains modèles d'apprentissage automatique, peuvent échouer. A cet effet, une corrélation des caractéristiques cytologiques, a été analysée en utilisant la matrice de Pearson [12]. Dans ce cadre, la matrice des corrélations est tout simplement la matrice des coefficients de corrélation statistiques calculés, sur plusieurs variables prises deux à deux.

Il s'agit dans le cas de cette analyse, des coefficients de Pearson. C'est donc la matrice des variances-covariances de variables réduites. La matrice est évidemment symétrique et sa diagonale est constituée de 1 puisque la corrélation d'une variable avec elle-même est parfaite (figure 4.4). Donc, sa trace est égale au nombre de variables.

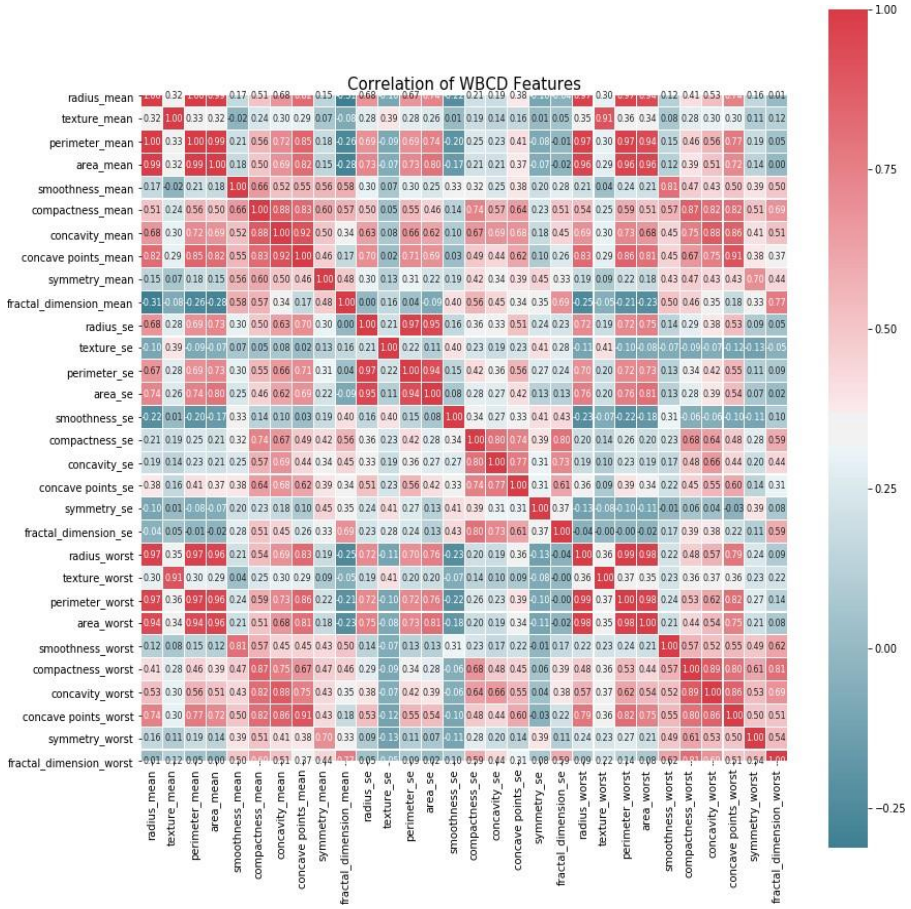


Figure 4.4 : Matrice de corrélation des caractéristiques cytologiques

Une analyse statistique est étudiée dans les figures 4.5 et 4.6, montrant ainsi les relations existantes entre ces paramètres. Les observations descriptives montrent que les valeurs moyennes, du rayon, du périmètre, de l'aire, de la compacité, de la concavité et des points concaves, peuvent positivement

être utilisés, dans la classification du cancer. En effet, les grandes valeurs de ces paramètres, montrent une corrélation avec les tumeurs malignes.

Cependant, les valeurs moyennes, de la symétrie, de la dimension fractale, de la texture et de l'uniformité, ne montrent aucune particularité dans le diagnostic. En effet, l'analyse statistique, ne montre pas de valeurs élevées de ces paramètres, qui nécessitent d'autres observations.

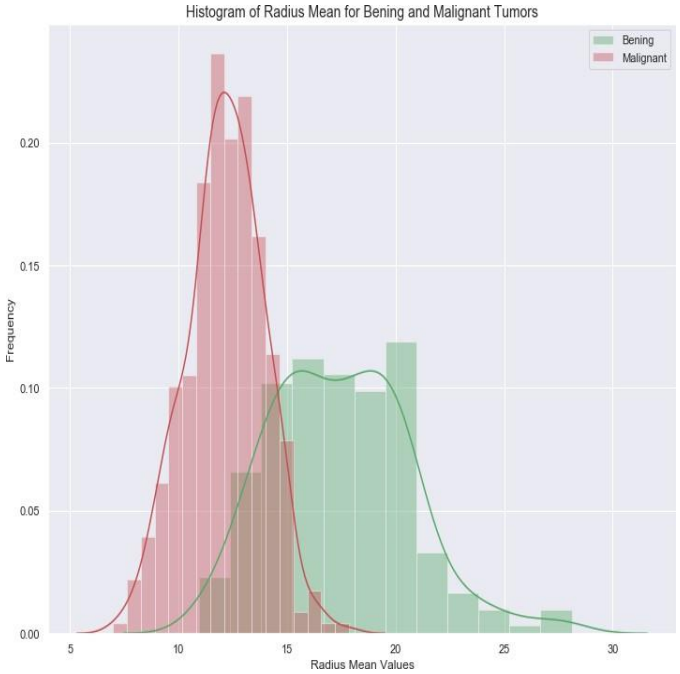


Figure 4.5 : Histogramme du rayon moyen des données cytologiques

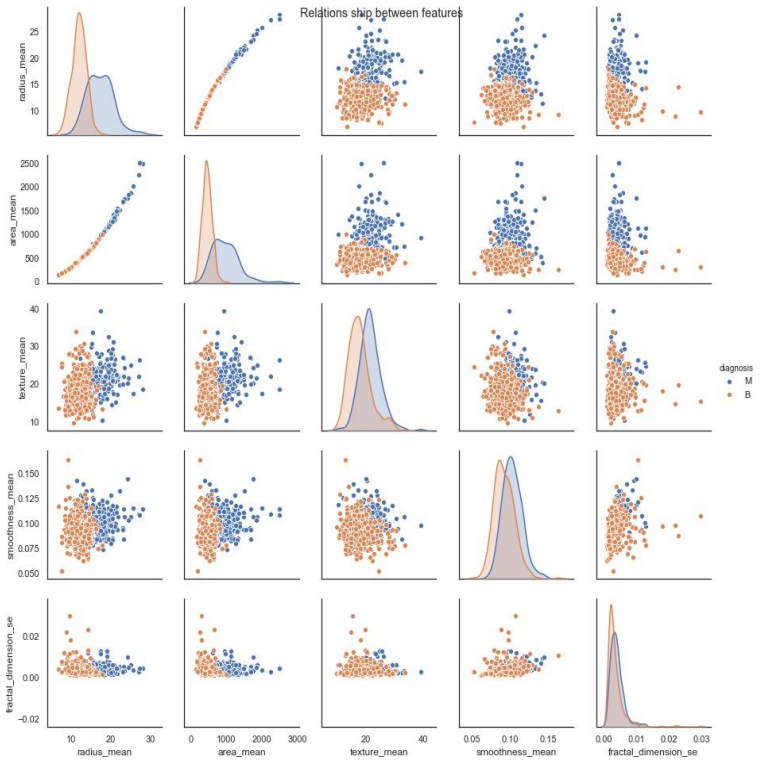


Figure 4.6 : Relations entre les caractéristiques cytologiques

La base de données présentée, doit être transformée en une matrice, pour effectuer l'apprentissage. Nous obtenons à cet effet dans la figure 4.6, où les colonnes représentent les caractéristiques et les lignes correspondent aux données.

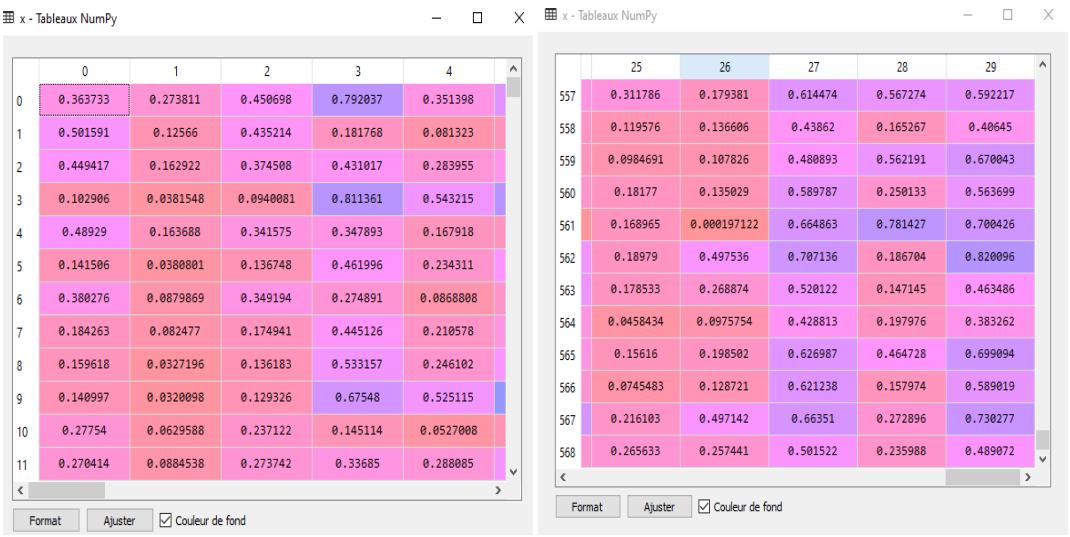


Figure 4.7 : Normalisation et conversion des données en une matrice

4.5 Résultats de BreastCytoLearn

BreastCytoLearn est un système qui a pour but de classifier les données cytologiques, en se basant sur les KNN, les SVM, les réseaux du perceptron multicouche et les convolutifs.

L'évaluation des performances d'un algorithme d'apprentissage automatique, est réalisée sur un ensemble d'entraînement et une base de tests. A cet effet, 70% de la base est consacrée à l'entraînement et 30% au test.

4.5.1 Classification 'BreastCytoLearn' par les K-NN

La première étape de ce classifieur, est de fixer le nombre de voisins K, pour calculer toutes les distances d'une observation X, avec les autres observations du jeu de données d'apprentissage.

La seconde étape est de détecter les k-voisins les plus proches des données du test à classer.

La dernière étape est d'attribuer les classes correspondantes, par vote majoritaire.

Lors de l'implémentation, nous calculons le taux d'erreur pour chaque, valeur de K, pour retenir le paramètre qui, minimise le taux d'erreur du test.

Les paramètres du classifieur sont représentés dans les figures 4.7, 4.8 et le tableau 4.1.

Base de données	Classifieur utilisé	Nombre des voisins optimums	Meilleur choix de K	Précision	Erreur
Wisconsin	K-NN	K = 13	K= 11	98%	0.08

Tableau 4.1 : Paramètres et précision du classifieur K-NN

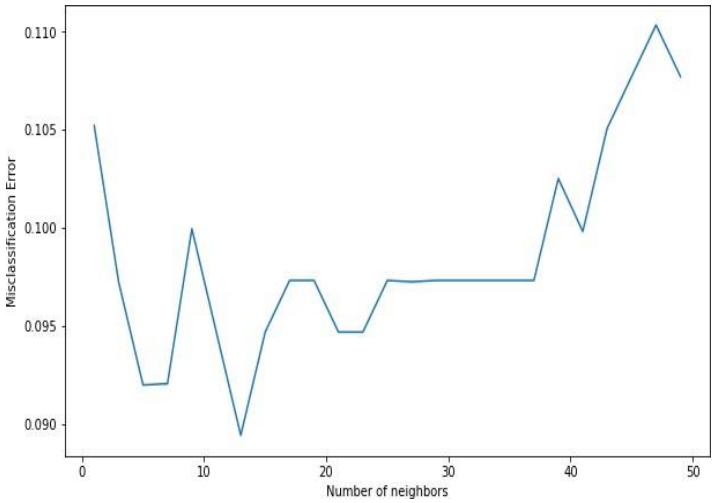


Figure 4.8 : Taux d'erreur de la classification en fonction de K

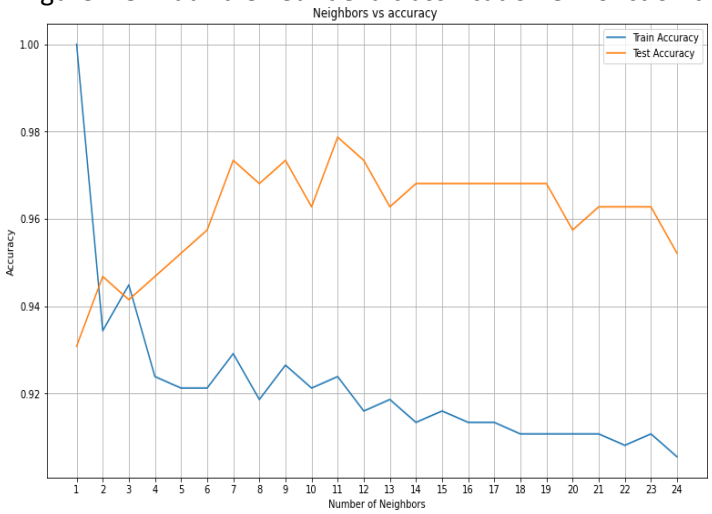


Figure 4.9 : Apprentissage et test des K-NN

4.5.2 Classification des données cytologiques par les SVM

L'objectif des SVM, est de maximiser la marge selon la relation vue dans le chapitre 2.

L'apprentissage, se fait, suivant différents kernels, pour la séparation des classes.

Les SVM ont engendré de bons résultats, avec un noyau radial régularisé (tableau 4.2). Les meilleures valeurs des paramètres de régularisation C= 2.295378978632264, et Gamma = 0.025915067554470612. C a été fixé en étudiant différentes valeurs (figure 4.9). L'avantage des SVM, est la rapidité de l'entraînement et l'optimisation des données.

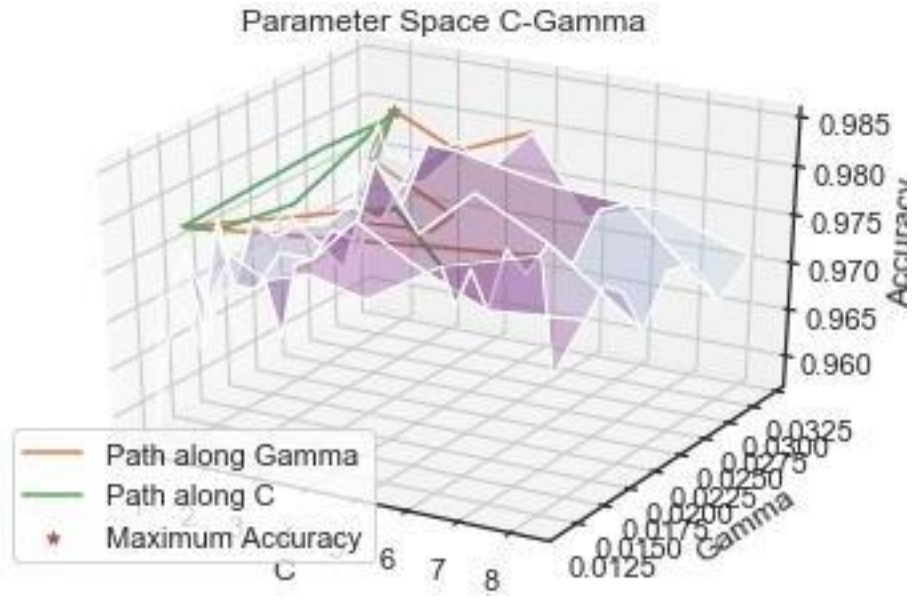


Figure 4.10 : Précision des SVM en fonction de C et Gamma

Bases de données	Classifieur utilisé	Noyaux utilisées	Meilleur noyau	Précision
Wisconsin	SVM	▪ Radial ▪ Linéaire ▪ Sigmoïde ▪ Gaussien	Radial	99%

Tableau 4.2 : Paramètres de la classification SVM

	Masse maligne	**Masse bénigne**
Masse maligne	73	2
Masse bénigne	0	113

Tableau 4.3 : Matrice de confusion pour les tests

Sur l'ensemble des données des tests, sur 75 cas malins, 73 ont été reconnus vrais positifs et 2 ont été confus avec des lésions bénignes.

Sur les 113 classes bénignes, toutes les lésions ont été reconnues à 100%.

4.5.3 Classification par le perceptron multicouche

Le perceptron multicouche étudié dans ce cas, adopte une couche d'entrée, qui reçoit les vecteurs caractéristiques des données cytologiques, deux couches cachées et une coche de sortie, pour représenter les classes. L'activation des couches, se fait par la fonction sigmoïde.

L'optimisation est effectuée par rétropropagation du gradient sur l'ensemble du réseau en utilisant la librairie tensorflow. Le pas du gradient est mis à jour après chaque passage sur l'ensemble des données, selon la formule learningRate = learningRate(1 − δ) où δ désigne la différence entre les taux de bonne classification aux itérations t − 1 et t. Le pas de mise à jour initial, est réglé à 10^{-2}.

La figure ci-dessous, représente la précision en fonction de 1000 itérations, lors de l'entrainement. A cet effet, la précision obtenue lors de l'entraînement, est de 98.24%.

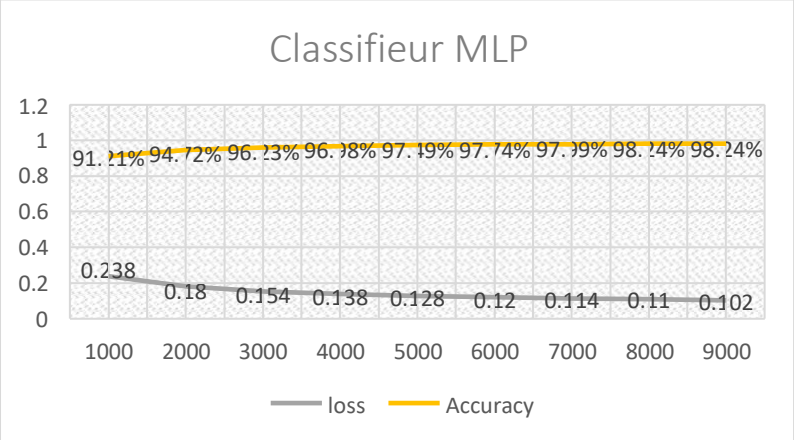
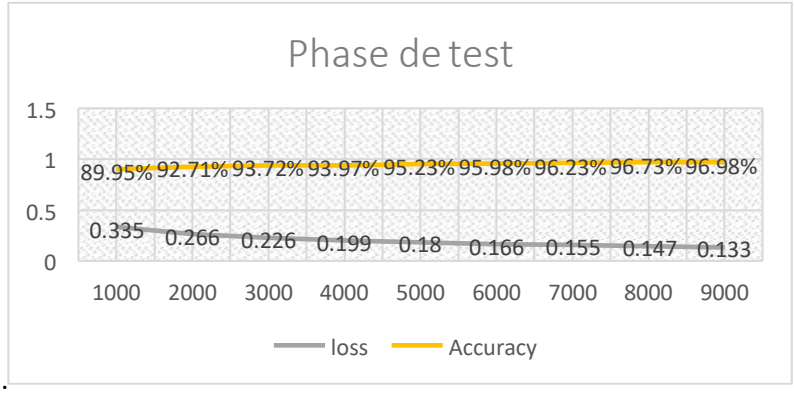


Figure 4.11 : Entraînement du MLP



.

Figure 4.12 : Phase de test

Lors de la phase des tests, la précision du classifieur MLP, est de 97%

4.5.4 Classification des données cytologiques par les CNN

Le système 'BreastCytoLearn' classifie les données cytologiques, en se basant aussi sur les CNN, à partir d'un réseau pré-entraîné. Comme les modèles utilisés, réalisent le traitement sur des données bidimensionnelles, une conversion a été faite, comme explicité dans la section 4.4.2.

Une représentation schématique des opérations faites sur le modèle du CNN basé sur le VGG16 (figure 4.12), est tracée sur la figure 4.13.

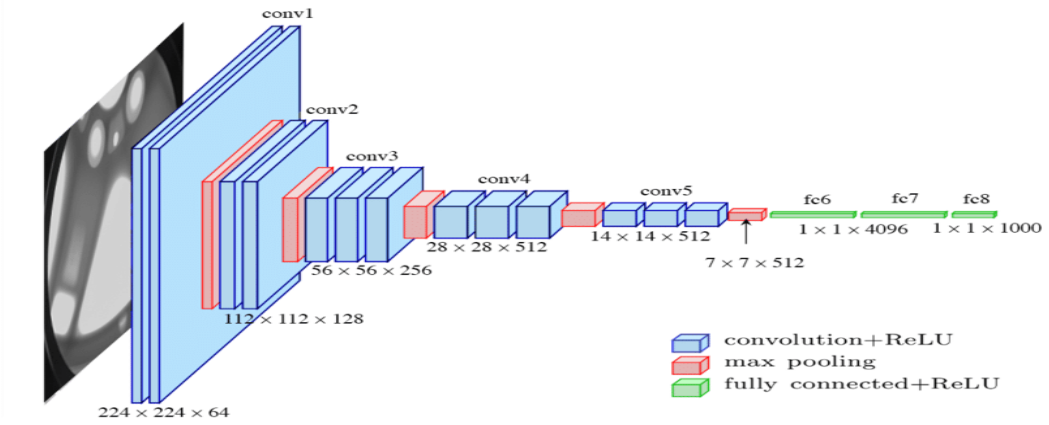


Figure 4.13 : Synoptique du VGG16

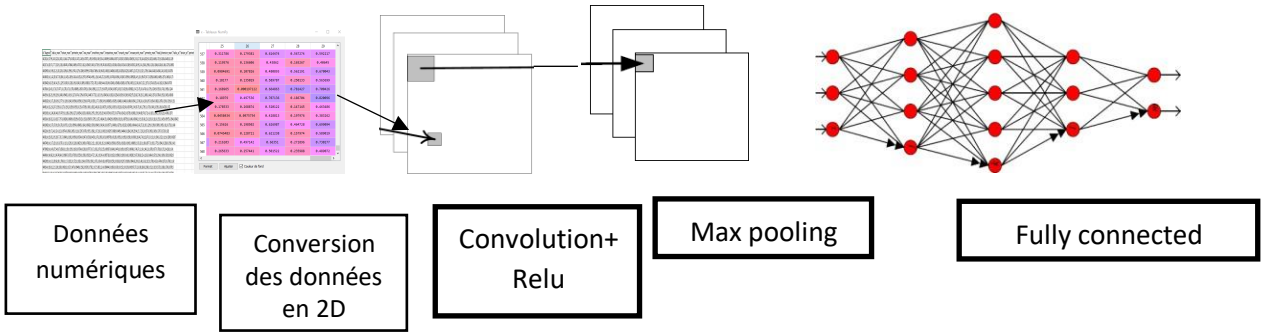


Figure 4.14: Synoptique de l'apprentissage par transfert des données cytologiques

Le modèle des CNN, s'effectue en deux étapes ; la première consiste en l'extraction des caractéristiques, suivant les opérations de convolution au nombre de 4 dans notre cas, suivies d'une rectification linéaire, par la fonction Relu, qui consiste à remplacer les valeurs négatives, par des 0. Une réduction des données est réalisée, par le max pooling, pour une sélection des caractéristiques.

La seconde étape, est la classification, qui s'effectue au niveau de la couche entièrement connectée, où les classes d'imagenet, sont remplacées par les deux classes cytologiques.

L'entraînement du réseau a été effectué par la méthode ADAM, la descente du gradient stochastique. A cet effet, nous avons réalisé la classification suivant trois cas. Le taux d'apprentissage est fixé à 0.01.

- **Premier Cas : Grand nombre d'itérations et un sur-apprentissage**

Les résultats provenant de ces expérimentations, sont regroupés dans le tableau 4.4.

Optimiseur	**Nombre d'itérations**	**Précision du modèle d'entrainement**	**Précision du modèle test**
Adam	600	100%	98%

Tableau 4.4 : Résultats du premier CNN

La fonction perte permet de quantifier l'écart entre les prévisions du modèle et les observations réelles du jeu de données, utilisé pendant l'apprentissage (figure 4.14).

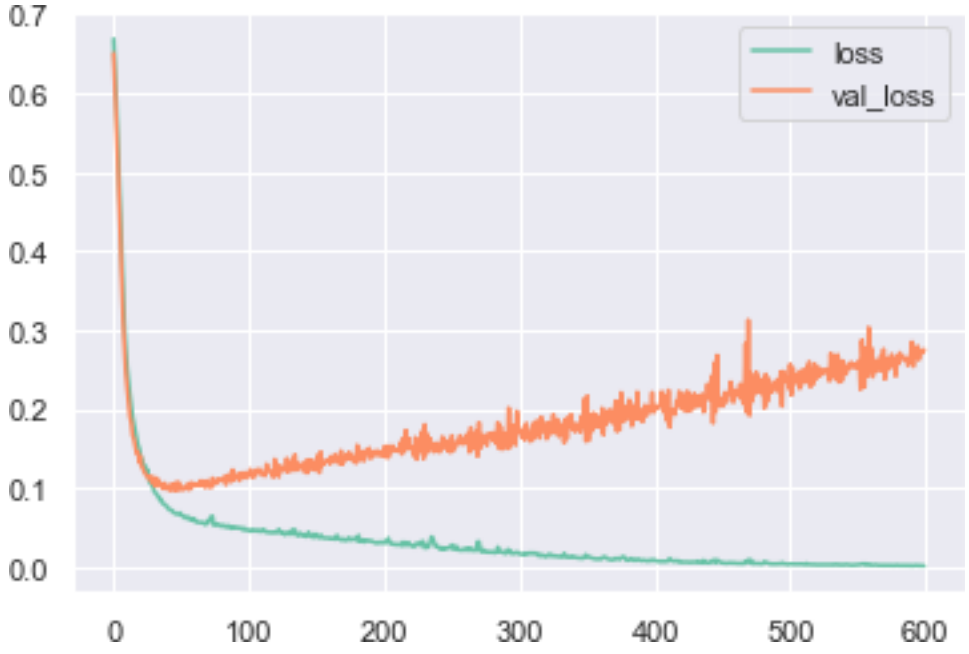


Figure 4.15 : Apprentissage et validation de la perte pour le 1er cas

L'entraînement est effectué avec une précision de 99%.

- **Deuxième cas : Arrêt de l'entrainement plus tôt**

Optimiseur	Nombre d'itérations	Nombre d'itérations avant l'arrêt de l'entrainement	Précision du modèle d'entrainement	Précision du modèle test
Adam	600	25	99%	97%

Tableau 4.5 : paramètres et précision du 2ème cas pour le classifieur CNN

Dans ce cas, lorsque l'arrêt de l'entraînement, se fait avant la totalité des itérations, la précision des tests, est légèrement améliorée et, l'erreur est stabilisée, relativement au premier cas.

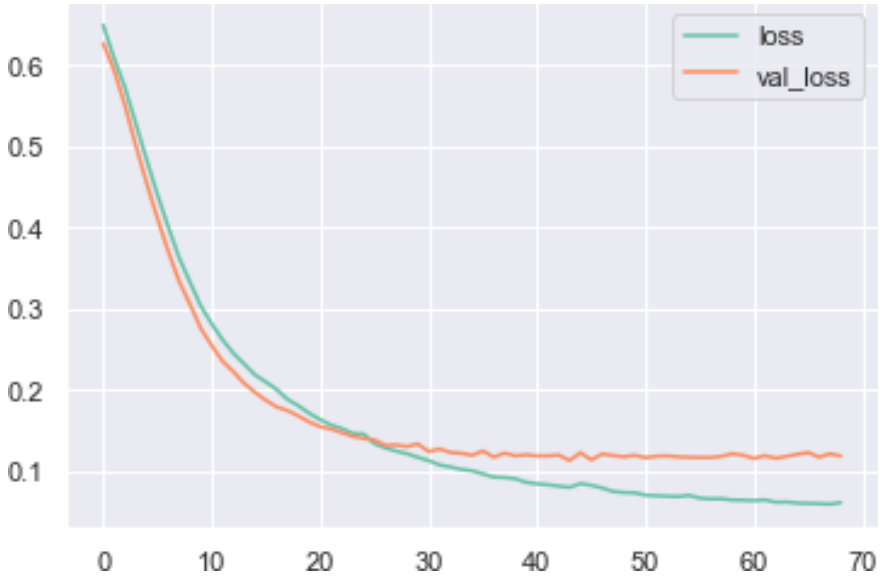


Figure 4.16 : Apprentissage et validation de la perte pour 2ème cas

- **Troisième cas : Avec ajout de couches Dropout**

Dropout est une technique de régularisation (pour combattre l'overfitting) dont le principe, est de désactiver aléatoirement à chaque itération, un certain pourcentage des neurones d'une couche. Cela évite ainsi le sur-apprentissage. Les résultats obtenus, sont regroupés dans le tableau 4.7.

Optimiseur	Nombre d'itérations	Précision du modèle d'entrainement	Précision du modèle test
Adam	600	99%	98%

Tableau 4.6 : paramètres et précision du 3ème cas des CNN

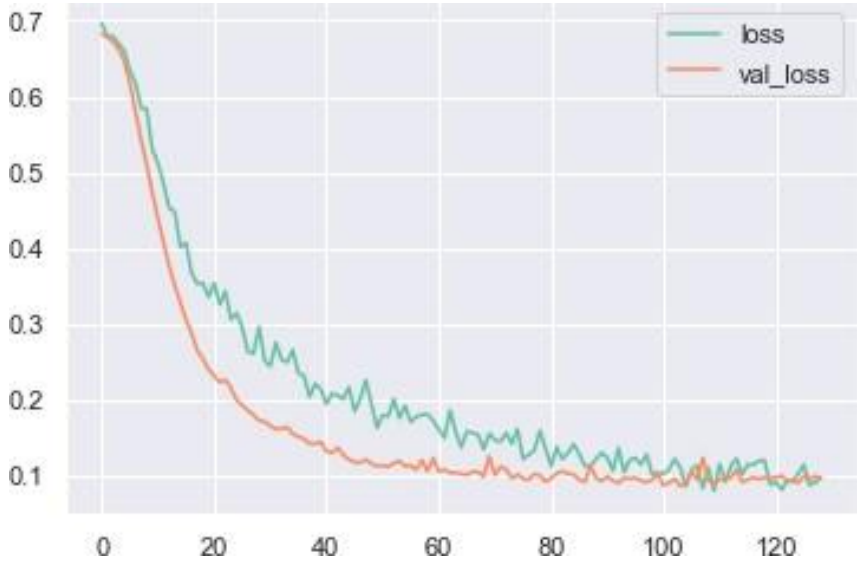


Figure 4.17 : Apprentissage et validation de la perte pour le 3ème cas

L'erreur s'est stabilisée à 0.1.

	Masse maligne	**Masse bénigne**
Masse maligne	56	1
Masse bénigne	0	87

Tableau 4.7: Matrice de confusion des tests

Selon la matrice de confusion, le modèle reconnaît 56 lésions malignes sur 57. La totalité des masses bénignes, est reconnue. La précision des tests, est alors de 98%.

4.6 Discussion

Les résultats obtenus, par les quatre classifieurs sont encourageants, puisque la précision obtenue, dépasse largement les 96%.

L'étude a été faite sur une base cytologique caractérisée et classifiée en deux catégories : les classes malignes et bénignes.

Les expérimentations ont été réalisées avec une variation des hyperparamètres, tels que le paramètre de pénalité C dans les SVM, pour réguler la classification, le taux d'apprentissage, pour contrôler les fonctions d'optimisation et le nombre d'itérations, pour suivre la phase d'entraînement.

Un petit récapitulatif (figure 4.17) nous montre les résultats obtenus pour les quatre classifieurs :

- le K-NN indique une précision de 98%.
- Les SVM génèrent une précision de 99%.
- Le MLP donne une précision de 97%.

Le modèle des CNN adopté dans trois cas différents, engendrent les précisions de :

- Le 1er cas indique une précision de 95%.
- le 2ème cas montre une précision de 97%.
- le 3ème cas permet une précision de 98%.

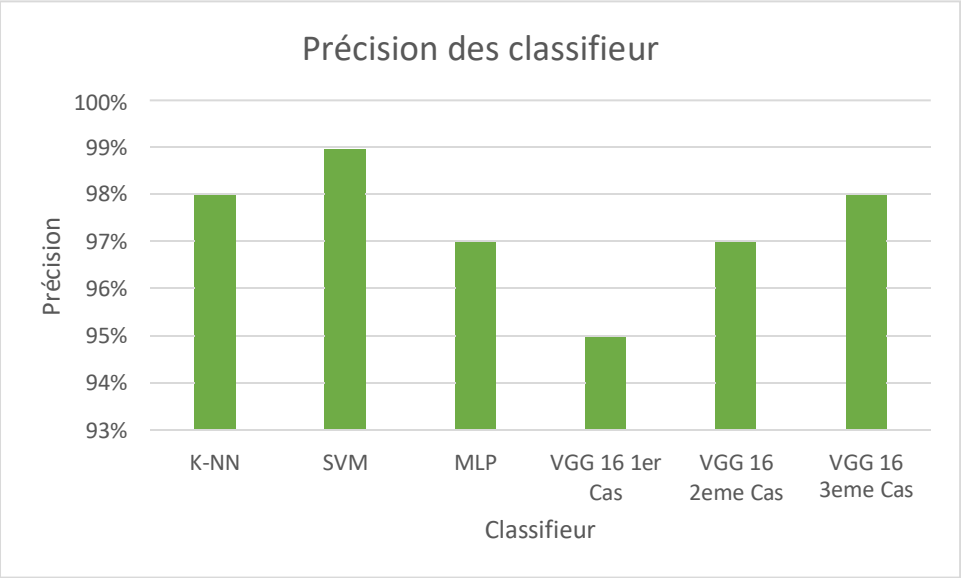


Figure 4.18 : Graphe représentant la précision obtenue pour chaque classifieur

Nous remarquons que le classifieur des SVM a engendré la meilleure précision

4.7 Conclusion

Dans ce chapitre, nous avons développé un système dédié à l'analyse des données cytologiques, par leur classification suivant les approches de l'intelligence artificielle. En tenant compte du choix des hyperparamètres, pour la construction des différents modèles, BreastCytoLearn a abouti à des résultats satisfaisants.

Conclusion générale

L'initiation à la recherche réalisée dans le cadre de ce mémoire, s'inscrit dans les objectifs de l'équipe, ''IMAD'', du laboratoire LATSI, pour la conception de systèmes d'aide au diagnostic en imagerie médicale.

Dans ce projet, nous avons étudié et développé, un système de classification morphologique des tumeurs mammaires, en vue d'une aide à la décision. Nous nous sommes focalisés sur les travaux rencontrés dans la littérature et les travaux du laboratoire LATSI, afin d'opter pour les méthodes d'apprentissage automatique et profond.

Nous avons concentré notre étude, en premier sur l'analyse statistique des données, pour voir l'effet des caractéristiques sur le type des lésions. En effet, certaines caractéristiques, comme la compacité, sont plus dominantes que d'autres. Ce qui introduit une première discussion, sur l'efficacité ou non de certains paramètres.

Le système réalisé, sur la base de l'analyse descriptive des données cytologiques du Wisconsin, s'est illustré par le développement de quatre classifieurs, à la recherche du meilleur pour le cas étudié.

Nous avons implémenté quatre classifieurs différents pour catégoriser les pathologies cytologiques mammaires. Trois modèles ont été implantés pour l'apprentissage automatique : les K plus proches voisins, les SVM et le MLP, le perceptron multicouche. Un quatrième modèle a suivi les premières approches. Il se base sur les réseaux convolutifs profonds, en tenant de trois cas différents. Le modèle en lui-même, est basé sur le réseau pré-entraîné VGG 16, suivant les modifications de certaines couches, telles que la couche entièrement connectée.

L'implémentation a été faite avec le langage de programmation python, en utilisant les bibliothèques adéquates pour faciliter la création de 'BreastcytoLearn' et pour l'accélération de l'entraînement.

Les résultats obtenus, suivant des hyperparamètres choisis selon nos expérimentations, ont engendré une précision qui dépasse les 96%. Nous pouvons conclure que le système BreastCytoLearn, pourrait aider le cytopathologiste dans sa prise de décision.

Les perspectives possibles, pour l'amélioration de ce travail, seraient de tester d'autres modèles des CNN, tels que Densenet, Resnet, ou les hybrider.

Il serait aussi intéressant, de considérer d'autres bases de données cytologiques présentant les différents grades des lésions cancéreuses mammaires, pour réaliser une approche globale d'aide à la décision.

Ce travail nous a permis d'acquérir énormément de connaissances, sur le 'machine learning' et le 'deep Learning'. Le temps passé à lire des livres et des articles, nous a servi à nous initier à la recherche.

Bibliographie

[1] Organisation mondiale de la santé. Cancer du sein : prévention et lutte contre la maladie, https://www.who.int/topics/cancer/breastcancer/fr/index3.html, date de consultation, novembre 2019.

[2] Santé Magazine. Cancer du sein, https://www.santemagazine.fr/sante/maladies/cancer/cancer-du-sein/cancer-du-sein-depistage-traitements-200412 , date de consultation, novembre 2019.

[3] M. Ozanam. Anatomie du sein, https://www.doctissimo.fr/html/sante/mag_2002/sem02/mag1004/dossier/sa_5966_anatomie_sein.htm , septembre 2016, date de consultation, novembre 2019.

[4] Société Canadienne du Cancer. Symptômes du cancer du sein, https://www.cancer.ca/fr-ca/cancer-information/cancer-type/breast/signs-and-symptoms/?region=on, date de consultation, 23 novembre 2019.

[5] Maxime Lambert. Cancer du sein : dépistage, symptômes, traitement, causes, où en est-on ?, https://www.maxisciences.com/cancer/cancer-du-sein-depistage-symptomes-traitement-causes-ou-en-est-on_art35205.html, décembre 2018, date de consultation, décembre 2019.

[6] Passeport Santé. Le cancer du sein, https://www.passeportsante.net/fr/Maux/Problemes/Fiche.aspx?doc=cancer_sein_pm, date de consultation, janvier 2020.

[7] Think Pink. Les sortes de cancer, https://www.think-pink.be/fr/articles/d/a/78/Les-sortes-de-cancer-du-sein, janvier 2013, date de consultation, janvier 2020.

[8] Fondation contre le cancer. La mammographie, https://www.cancer.be/le-cancer/jeunes-et-cancer/les-examens/la-mammographie, date de consultation, janvier 2020.

[9] Docteur Benchimol. La tomosynthèse mammaire ou mammographie 3D, 26 décembre 2017 mis à jour le 9 janvier 2018, https://www.docteur-benchimol.com/la-tomosynthese-mammaire-ou-mammographie-3d.html, date de consultation, janvier 2020.

[10] F. Colombier. Densité mammaire : ce qu'il faut savoir, https://www.doctissimo.fr/sante/cancer-sein/prevention-du-cancer-du-sein/densite-mammaire, octobre 2019, date de consultation, janvier 2020.

[11] Mammographie, disponible sur https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Diagnostic/Mammographie, date de avril 2020.

[12] I. Cheikhrouhoue, Description et classification des masses mammaires pour le diagnostic du cancer du sein, thèse de doctorat en informatique, Université d'Evry-Val d'Essonne, 2012.

[13] Classification de Le Gal des microcalcifications mammaires, http://www.aly-abbara.com/echographie/biometrie/scores/microcalcification_classification_le_gal.html, date de consultation, janvier 2020.

[14] W. Benhenia, F. Djillali. Analyse texturale de cellules tumorales mammaires pour l'aide à l'interprétation, mémoire de Master en Informatique, Université Sâad Dahlab-Blida 1, Blida, juillet 2012.

[15] Institut National du Cancer. Ponction cytologique, https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Diagnostic/Ponction-cytologique, date de consultation, juin 2020.

[16] M.A Kara Zaitri. Prélèvement cytologique, http://www.dr-karazaitri-ma.com/pages/examens-complementaires/autres-explorations-du-sein/cytoponction-1.html, date de consultation, en juin 2020.

[17] R.K, « Machine Learning » : Apprentissage supervisé ou non supervisé, https://medium.com/@bobkrc/machine-learning-apprentissage-supervisé-ou-non-supervisé-bced5be4fd7f, mars 2018, date de consultation, janvier 2020.

[18] IBM, le Machine Learning et la science des données, https://www.ibm.com/fr-fr/analytics/machine-learning

[19] M. K. Moudachirou. Classification et forêts aléatoires : Application à l'aide à la décision chirurgicale du genou par arthroplastie, mémoire de maîtrise en technologie de l'information, Télé-Université, Québec, Canada, juillet 2017.

[20] P. Gambette. Classification supervisée et non supervisée, Cours en ingénierie linguistique, Master 1 Informatique, Université Marne-la-Vallée (IGM), 2014.

[21] B. Paarth. Image Classification with K Nearest Neighbours, https://medium.com/swlh/image-classification-with-k-nearest-neighbours-51b3a289280, juillet 2019, date de consultation juin 2020.

[22] Y. Benzaki. Introduction à l'algorithme K Nearst Neighbors (K-NN), https://mrmint.fr/introduction-k-nearest-neighbors, 2018, date de consultation, juin 2020.

[23] V. Vapnik. The Nature of Statistical Learning Theory, N-Y, Springer-Verlag,1995

[24] S. Benyahia Belaidi. Application de la classe des méthodes d'apprentissage statistique SVM (support vector machine) pour la reconnaissance des formes dans les images Mémoire cadre de de Magister en Informatique, Université Abou Bakr Belkaid– Tlemcen, 2012.

[25] A. L. Gonzalez. Exploration des arbres de décision et des support vector machines en vue d'applications dans l'analyse de texte, maîtrise en mathématiques et informatique appliquées, Université du Québec à Trois-Rivières, 2016.

[26] C. ZHU. Effective and Efficient Visual Description based on Local Binary Patterns and Gradient Distribution for Object Recognition. Thèse de doctorat en informatique, Ecole centrale de Lyon 2012.

[27] G. Petitjean. Introduction aux réseaux de neurones, https://www.lrde.epita.fr/~sigoure/cours_ReseauxNeurones.pdf, date de consultation, juin 2020.

[28] Juri'Predis, le blog. Démystifier le Machine Learning, Partie 2 : les Réseaux de Neurones artificiels, https://www.juripredis.com/fr/blog/id-19-demystifier-le-machine-learning-partie-2-les-reseaux-de-neurones-artificiels, date de consultation, juin 2020.

[29] R. Lambert. Différence entre Intelligence Artificielle, Machine Learning et Deep Learning, décembre 2018, https://penseeartificielle.fr, date de consultation juin 2020.

[30] J. PARIS. L'intelligence artificielle : présentation, applications et limites, https://devotics.fr/presentation-intelligence-artificielle, septembre 2019, date de consultation, juin 2020.

[31] C. C. Aggarwal, Neural Networks and Deep Learning, Springer, 2018.

[32] : Ali Labiad, Sélection des mots clés basée sur la classification et l'extraction des règles d'association, Univérsité du Québec à Trois-Rivières, juin 2017.

[33] Data Analytic Post, Deep Learning, https://dataanalyticspost.com/Lexique/deep-learning/, juin 2020.

[34] N. A. D. DIALLO, La reconnaissance des expressions faciales, mémoire de Master en informatique, Université 8 Mai 1945, Guelma, juillet 2019

[35] : Charles Crouspeyre, Comment les Réseaux de neurones à convolution fonctionnent, juillet 2017, https://medium.com/@CharlesCrouspeyre/comment-les-r%C3%A9seaux-de-neurones-%C3%A0-convolution-fonctionnent-b288519dbcf8, juillet 2020.

[36] : Siddharth Das et Analytics Vidhya, CNN Architectures : LeNet, AlexNet, VGG,GoogLeNet, RestNet and more, 2017, https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5#:~:text=Top%20highlight-,CNN%20Architectures%3A%20LeNet%2C%20AlexNet%2C%20VGG,%2C%20GoogLeNet%2C%20ResNet%20and%20more%E2%80%A6&text=A%20Convolutional%20Neural%20Network%20(CNN,pixel%20images%20with%20minimal%20preprocessing , juillet 2020.

[37] Mugahed A. Al-antaria, Mohammed A. Al-masnia, Mun-Taek Choib, Seung-Moo Hana, Tae-Seong Kima, A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification, 2018

[38]: Zhongyi Han, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li & Shuo Li, Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model, 2017

[39] Anuraganand Sharma, Dinesh Kumar, Classification with 2-D Convolutional Neural Networks, for breast cancer diagnosis, arXiv:2007.03218, 26 pages, date de consultation juillet 2020.

[40] Claire D. Costa. Best Python Libraries for Machine Learning and Deep Learning, https://towardsdatascience.com/best-python-libraries-for-machine-learning-and-deep-learning-b0bd40c7e8c, date de consultation, date de consultation, juin 2020.

[41] R. Carver. Bibliothèques Python parmi les meilleures pour la finance, , https://news.efinancialcareers.com/fr-fr/3000290/voici-six-bibliotheques-python-parmi-les-meilleures-pour-la-finance, février 2019, date de consultation, juin 2020.

[42] Anaconda. Your data science toolkit, https://www.anaconda.com/products/individual, consulté en juin 2020.

[43] Spyder, https://www.spyder-ide.org/, consulté en juin 2020.

[44] S. Bancal. Introduction à Numpy, https://enacit.epfl.ch/cours/python/scientifique/numpy.html#numpy, consulté en janvier 2020.

[45] Binning in python and pandas, https://python.course.eu, date de consultation, juin 2020.

[46] Scipy stacks, https://scipy.org, date de consultation, juin 2020.

[47] Scikit-learn, machine learning in python, https://scipy.org, date de consultation, juin 2020.

[48] EDUCBA. Introduction to Tensorflow, https://www.educba.com/introduction-to-tensorflow, date de consultation, janvier 2020.

[49] R. Vickery. Beginners Guide to Deep Learning with TensorFlow, https://towardsdatascience.com/beginners-guide-to-deep-learning-with-tensorflow-ca85969b2f2, date de consultation, en Janvier 2020.

[50] Keras, the python deep learning API, https://keras.io, date de consultation, juin 2020.

[51] Seaborn. Statistical data visualization https://seaborn.pydata.org, date de consultation, juin 2020.

[52] W. H. Wolberg. Breast cancer Wisconsin data set, http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29, date de consultation, date de consultation, avril 2020.