République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université ABDERRAHMANE MIRA - BEJAIA -

Faculté Science Exacte
Département Mathématiques Appliquées

Mémoire pour l'obtention du diplôme de master en mathématiques appliquées
Option : Sciences de données et aide à la décision

Analyse et Classification de Données pour la Prédiction et l'Aide au Diagnostic du Cancer du Sein

Élaboré par :
- OUAZENE Amazigh
- BACHOUCHE Celine

Encadré par :
- Dr BOUZIDI Lhadi

Devant les jurys :
- K. KRIMAT
- N. ZOUGAB
- L. HARFOUCHE

2023-2024

Remerciment

Nous remercions Dieu tout-puissant d'avoir guidé nos pas vers les portes du savoir tout en illuminant nos chemins et de nous avoir donné suffisamment de courage, de force et de persévérance pour mener notre travail à terme.

Nous tenons à remercier M. Bouzidi Lhadi, qui nous a permis de bénéficier de son encadrement, s'est toujours montré à l'écoute et disponible tout au long de la réalisation de ce mémoire, ainsi que pour l'inspiration, la patience, l'aide et le temps qu'il a bien voulu nous consacrer durant l'accomplissement de ce mémoire.

L'aboutissement de ce travail nous donne l'occasion d'exprimer notre sincère reconnaissance au Dr Kadi Yacine, notre maître de stage, pour sa confiance et les connaissances qu'il a su partager avec nous. Nous le remercions aussi pour sa disponibilité et la qualité de son encadrement dans son cabinet d'imagerie et de radiologie, ainsi que tout le personnel y travaillant.

Nos remerciements s'adressent de même à tous nos enseignants de la Faculté des Sciences Exactes, qui ont contribué à notre formation durant ces cinq années.

Enfin, nous remercions toutes les personnes ayant contribué de près ou de loin à l'élaboration de ce modeste travail.

Dédicace

Je dédie ce modeste travail à tous ceux qui m'ont soutenu tout au long de mon parcours d'études.

À mes très chers parents, maman et papa, qui sont la source de ma réussite. Je n'oublierai jamais non plus mes sœurs, toi Lina et Anias, que Dieu vous garde pour moi.

À mes grands-parents, ma chère grand-mère Fatma, mon précieux grand-père Hassan, à mon défunt grand-père « Mahfoud » que le bon Dieu lui garde une place dans son vaste paradis.

Sans oublier de faire un saut chez mes chères tantes maternelles, mes oncles Yacine et Mahmoud, mes chers cousins et cousines, je vous souhaite la réussite.

À mes chers amis Amazigh, Katia, Boualem,Lucie, je vous souhaite tout le bonheur et la réussite.

Sans oublier mon binôme « Amazigh » qui m'a aidé et soutenu. Je lui souhaite une vie remplie de bonheur, de réussite professionnelle et l'obtention d'un excellent poste de travail.

Finalement, je dédie ce mémoire à tous ceux qui m'aiment du fond du cœur, surtout ceux qui m'ont soutenu dans les moments difficiles. Merci pour tout, et à toutes les personnes qui m'ont aidé, même avec un beau mot ou un beau geste.

Dédicace

C'est avec une très grande émotion que je dédie ce modeste travail à mes très chers parents « Saïd et Katia », la source de mon existence, qui m'ont accompagné durant tout mon parcours.

Je dédie à mes grands-parents, mes chères grand-mères Fatma et Zitouma, mon précieux grand-père Abdellah, à mon défunt grand-père « Mohamed » que le bon Dieu lui garde une place dans son vaste paradis.

À mon cher frère Billal et ma chère sœur Hannane, que Dieu vous garde pour moi, je vous souhaite la réussite.

Sans oublier de faire un saut chez mes tantes et oncles maternels et paternels, mes chers cousins et cousines, je vous souhaite le bonheur et la réussite. Ainsi qu'à l'équipe de Domino.

En fin de compte, je consacre ce mémoire à tous ceux qui m'aiment sincèrement, en particulier ceux qui m'ont soutenu pendant les moments difficiles. Je suis reconnaissant envers tous, y compris ceux qui m'ont aidé avec un mot ou un geste précieux.

Sans oublier mon binôme, Céline, qui m'a apporté son aide et son soutien précieux. Je lui souhaite une vie pleine de bonheur, une carrière réussie et l'obtention d'un poste de travail exceptionnel.

Table des matières

Table des figures

Liste des tableaux

Acronymes

TNM : Tumeur,Nodes,métastase
ACR : American College of Radiology
IA : Intelligence Artificielle
ML : Machine Learning
RNA : Réseaux de Neurones Artificiels
ANN : Artificial Neural Network
MLP : Multi-Layer Perceptron
RNN : Recurrent Neural Network
CNN : Convolutional Neural Network
DP : Deep Learning
TPU : Tensor Processing Unit
GPU : Graphics Processing Unit

Introduction générale

Dès son apparition, l'homme a été confronté à une multitude de maladies. Certaines ont disparu au fil du temps, tandis que d'autres persistent et gagnent en importance dans les domaines médico-social et médico-sanitaire. Parmi ces pathologies, le cancer se distingue comme une menace majeure pour la santé publique, nécessitant une intervention et une prise en charge urgentes. La lutte contre ce fléau exige une mobilisation accrue des ressources et des efforts de la communauté scientifique, médicale et politique.

Le cancer du sein est le cancer le plus fréquent chez les femmes dans le monde, et il représente environ 25 % de tous les cancers féminins. En 2020, on estime qu'environ 2,3 millions de nouveaux cas de cancer du sein seront diagnostiqués dans le monde, et que 685 000 femmes mourront de la maladie. Au cours des 5 dernières années, 7,8 millions de femmes ont été guéries du cancer du sein et ont survécu.il peut survenir chez les femmes de tout âge. Cependant, il est plus fréquent chez les personnes âgées [1]

Dans cette ère de progrès technologique rapide, l'intelligence artificielle (IA) émerge comme un catalyseur potentiel pour transformer la compréhension et la gestion du cancer du sein. Elle offre des capacités analytiques avancées, permettant une interprétation plus rapide et précise des données médicales, en particulier dans le domaine de l'imagerie diagnostique. Cette technologie présente ainsi l'opportunité de révolutionner les méthodes traditionnelles de diagnostic et de prédiction du cancer du sein. Loins de prétendre une étude approfondi de cette patologie, l'étude que nous menons ici est surtout réalisé dans le but d'identifier les différents concepts qui vont nous permettre de construire un modèle de machine learning et de deep learning d'aide au diagnostic et à la prédiction de cette maladie.

Problèmatique

Le diagnostic précoce du cancer du sein reste effectivement un défi majeur malgré les avancées scientifiques. Plusieurs facteurs peuvent compliquer ce processus :

① La variabilité des tumeurs et leur taille initiale limitée peuvent rendre leur détection difficile, particulièrement chez les femmes à seins denses.

② Les examens courants comme la mammographie et la biopsie ont des limites de sensibilité et de spécificité, pouvant manquer certaines tumeurs ou entraîner des faux positifs.

③ L'interprétation des images médicales dépend de l'expertise du radiologue et du pathologiste et peut être sujette à des erreurs humaines.

C'est là que l'Intelligence Artificielle (IA) peut apporter des solutions prometteuses :

① **Détection assistée par IA** : Les algorithmes d'IA formés sur de grandes quantités de données d'imagerie peuvent aider à détecter plus tôt et avec plus de précision les anomalies suspectes, même infimes.

② **Analyse d'images multimodales** : En combinant différentes modalités d'imagerie (mammographie, échographie, IRM, etc.), l'IA peut exploiter des informations complémentaires pour un diagnostic plus fiable.

③ **Aide à la décision clinique** : En intégrant non seulement les données d'imagerie mais aussi d'autres facteurs (antécédents, génétique, etc.), l'IA peut fournir aux médecins des recommandations personnalisées pour une prise en charge optimale.

Cependant, il est crucial que ces systèmes d'IA soient développés de manière rigoureuse, validés sur des données réelles et utilisés en complément de l'expertise médicale, et non comme un remplacement. Une approche combinant l'IA et les professionnels de santé offre le potentiel d'améliorer significativement le dépistage et le diagnostic précoce du cancer du sein, tout en rassurant les patientes.

Objectifs

Notre objectif est de mettre en place un système d'intelligence artificielle capable d'analyser efficacement les données médicales pertinentes pour le diagnostic du cancer du sein, en intégrant des techniques avancées d'apprentissage automatique et profond. Ce système vise à améliorer la précision et la fiabilité du diagnostic en identifiant de manière précoce la maladie. En outre, l'objectif est de développer une interface conviviale permettant aux professionnels de la santé d'utiliser facilement cette technologie dans leur pratique quotidienne, contribuant ainsi à améliorer les résultats cliniques et à réduire l'angoisse des patients face au diagnostic du cancer du sein.

Organisation du mémoire

Ce mémoire est organisé en 5 chapitre :

① **Chapitre I Généralités sur le cancer de sein :**
Nous abordons le concept du cancer du sein, ainsi que l'anatomie féminine du sein, les types de cancer du sein, les symptômes les plus courants, les facteurs de risque et les traitements du cancer du sein.

② **Chapitre II Apprentissage automatique et apprentissage profond :**
Nous avons examiné comment les méthodes d'apprentissage automatique et profond peuvent être utilisées pour analyser et classer les données médicales, mettant en évidence leur efficacité dans la prédiction et le diagnostic du cancer du sein.

③ **Chapitre III Etat de l'art :**
Il aborde les principales approches et techniques pour faciliter le diagnostic du cancer du sein, soulignant l'efficacité des algorithmes de Deep Learning pour établir un lien entre l'informatique et la médecine.

④ **Chapitre IV Conceptions et résultats :**
Il parle de la comparaison des algorithmes utilisés pour prédire le cancer du sein à partir de données tabulaires, avec l'ANN comme le modèle le plus précis, et le diagnostic basé sur l'imagerie, où le modèle ResNet50 s'est révélé le plus précis.

⑤ **Chapitre V Réalisation d'une application de déploiement**
Il parle de la création d'un outil de prise de décision pour la prédiction et la classification des tumeurs mammaires en utilisant Taipy, offrant une application interactive pour les professionnels de la santé, facilitant ainsi la prise de décisions cliniques avec des analyses précises.

ChapitreI

Généralités sur le cancer de sein

I.1 Introduction

Le cancer est l'un des plus mortels de l'histoire, infectant les femmes par un pourcentage important, et la deuxième cause de décès chez les femmes après un cancer de poumon.

Le cancer du sein peut également toucher les hommes ,ils concernent moins de 1 % des cancers du sein.

En Algérie, le cancer du sein est le cancer le plus fréquent chez les femmes, avec plus de 14 000 nouveaux cas diagnostiqués chaque année et que 4500 femmes mourront de la maladie. Il représente la première cause de mortalité par cancer chez les femmes algériennes. [2]

I.2 Définition

Le cancer du sein est une maladie caractérisée par la croissance incontrôlée de cellules anormales dans les tissus du sein. Ces cellules cancéreuses peuvent former une tumeur si elle n'est pas traitée, elle peut envahir les tissus environnants et se propagera a d'autres parties du corps. [1]

Il peut commencer dans les canaux qui transportent le lait vers le mamelon(cancer canalaire) ou dans les glandes productrices de lait (cancer lobulaire).

Deux principales catégories de tumeurs sont identifiées :

- **Tumeur bénigne**
- **Tumeur malignes**

I.2.1 Tumeur bénigne

Est une masse de cellules anormales qui se développe lentement et ne s'étend pas à d'autres parties du corps. Elle est généralement enlevée chirurgicalement pour des raisons esthétiques ou pour prévenir les complications.

La tumeur bénigne qui se développe le plus souvent dans le sein est le fibroadénome. Les autres affections bénignes du sein sont les :

- **Kystes**
- **Les changements fibrokystiques**
- **L'hyperplasie**
- **L'écoulement du mamelon**
- **La gynécomastie**

La plupart des masses du sein ne sont pas synonymes de cancer, mais seul l'examen anatomopathologique réalisé après une biopsie permet de vérifier qu'il ne s'agit pas d'un cancer.[3]

I.2.2 Tumeur malignes

Est une masse de cellules anormales qui se développe rapidement et peut s'étendre à d'autres parties du corps. Elle est également connue sous le nom de cancer. Les cellules cancéreuses qui composent les tumeurs malignes présentent diverses anomalies par rapport à des cellules normales (forme et taille différentes, contours irréguliers,...etc).
On parle de cellules indifférenciées car elles ont perdu leurs caractéristiques d'origine.

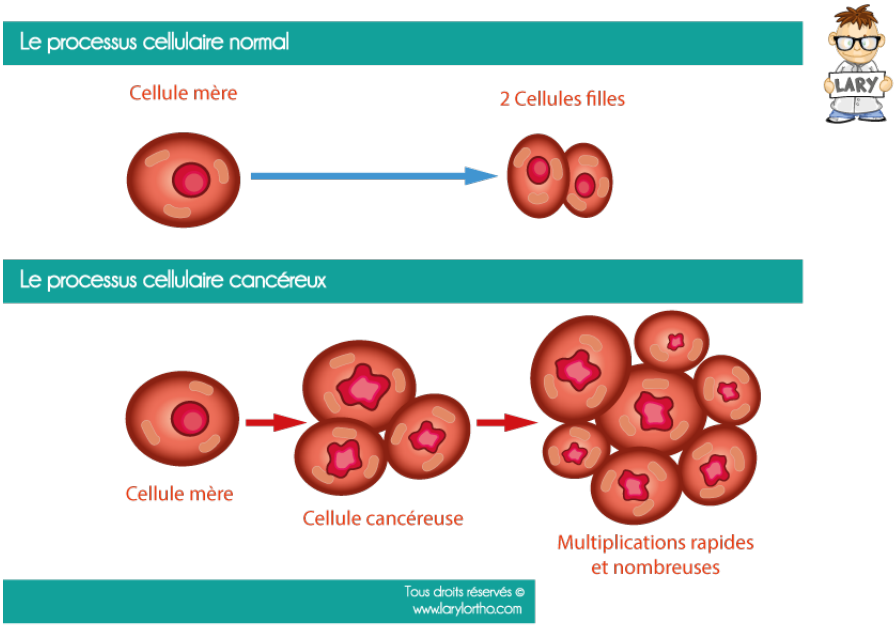


FIGURE I.1 – Division des cellules cancéreuses

I.2.3 Anatomie du sein

Le sein et un organe pair et symétrique de forme hémisphérique, situé en avant du thorax, entre la troisième et la cinquième cote au-dessus du muscle grande pectoral. La fonction biologique du sein est de produire du lait afin de nourrir un nouveau-né. Chaque sein contient une glande mammaire (elle -même composée de quinze à vingt compartiments séparés par du tissu graisseux) et du tissu de soutien qui contient des vaisseaux, des fibres et de la graisse. Chacun des compartiments de la glande mammaire est constitué de lobules et de canaux. Le rôle des lobules est de produire le lait en période d'allaitement. Les canaux transportent le lait vers le mamelon. [3]

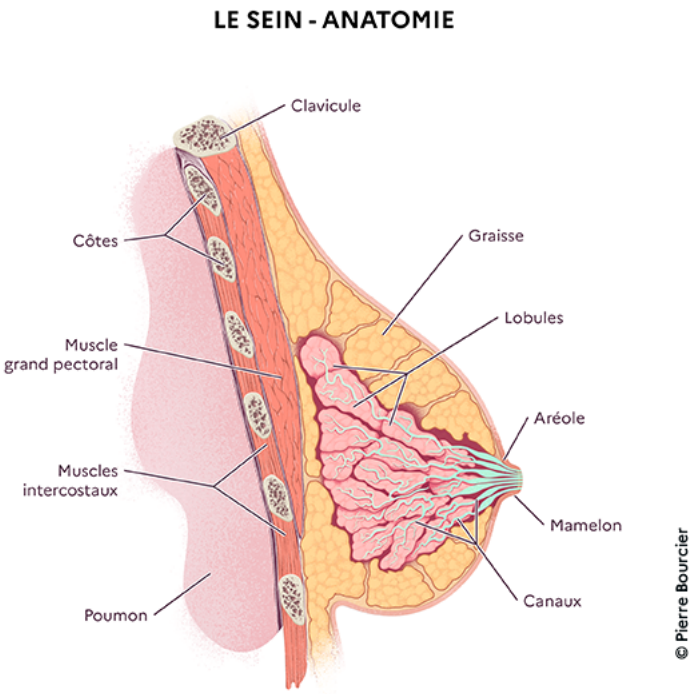


FIGURE I.2 – Anatomie du sein

I.3 Types du cancer

Carcinome canalaire in-situ : Ce type affecte les conduits qui transportent le lait de la glande mammaire au mamelon.

Carcinome canalaire infiltrant : Ce type atteint les conduits de lait, mais est plus invasif car il propage au tissu mammaire.

Carcinome lobulaire in-situ : Croissance de cellules anormales dans les glandes mammaires qui sécrètent du lait, mais ce de changement augmente le risque d'infection.

Carcinome lobulaire infiltrant : Ce type de cancer est plus rare, il débute dans les glandes mammaires puis se propage aux autres tissus du sein.

Cancer inflammatoire du sein : Il s'agit d'une forme rare de cancer de sein et agressif et semble se développer rapide [5].

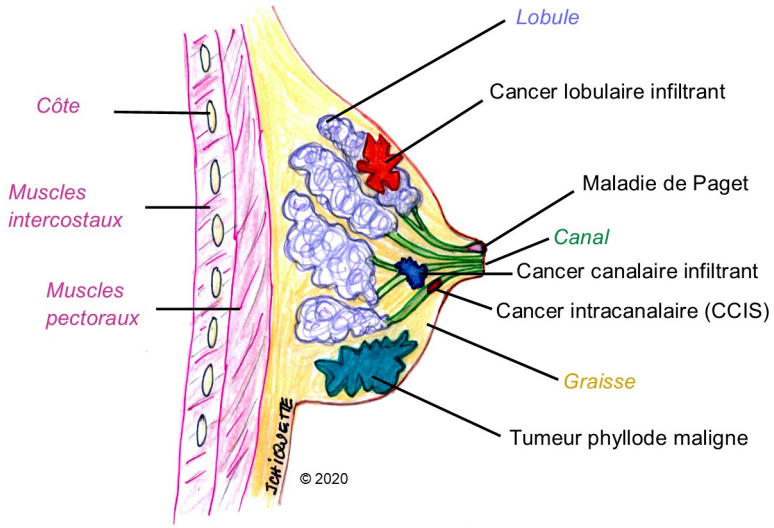


FIGURE I.3 – Les types du cancer du sein

I.4 Symptômes

Les symptômes du cancer du sein peuvent varier, mais certains signes doivent alerter [3] :

- ✦ Une masse ou une grosseur dans le sein.
- ✦ Changement de la forme ou de la taille du sein.
- ✦ Changement de la peau du sein, comme un creux ou un enflement.
- ✦ Changement du mamelon, comme un écoulement, une rétraction ou un changement de couleur.
- ✦ Douleur dans le sein ou l'aisselle.

I.5 Dépistage et diagnostic

Selon l'American Cancer Society, il est recommandé que les femmes âgées de 40 à 44 ans aient la possibilité de débuter le dépistage par mammographie chaque année. De 45 à 54 ans, elles devraient passer une mammographie chaque année. À partir de 55 ans, elles peuvent opter pour une mammographie tous les deux ans, ou continuer les examens annuels. Le dépistage doit se poursuivre aussi longtemps que la personne est en bonne santé et peut prévoir une espérance de vie d'au moins 10 ans.[4]

Il existe plusieurs méthodes pour diagnostiquer la pathologie du cancer du sein, notamment la mammographie, l'échographie mammaire, l'imagerie par résonance magnétique (IRM) mammaire, la biopsie, et l'examen clinique des seins.

- ✦ **La palpation :** Elle permet la mise en évidence d'une grosseur anormale.
- ✦ **L'imagerie :**
  - ✦ **La mammographie :** Est une radiographie à faible dose du sein. L'image obtenue par la mammographie est appelée cliché mammaire. Elle peut aider à

détecter des tumeurs cancéreuses (malignes) et des tumeurs non cancéreuses (bénignes) dans le sein. [5]

- ✦ **L'échographie :** Est un examen d'imagerie qui emploie les ondes sonores de haute fréquence pour produire des images d'organes et de structures du corps. Lors de l'échographie, des ondes sonores sont réfléchies par différentes parties du corps.[5]
- ✦ **L'imagerie par résonance magnétique (IRM) :** Est un examen d'imagerie qui emploie de puissantes forces magnétiques, des ondes radioélectriques et un ordinateur pour produire des images détaillées à 3 dimensions d'organes, d'os et de tissus mous à l'intérieur de votre corps. Certains examens d'IRM requièrent un produit de contraste [5].

✦ **Les prélèvements :**

- ✦ **La cytoponction :** Prélèvement de quelques cellules avec une aiguille très fine afin de les analyser.
- ✦ **La biopsie :** Consiste à prélever un échantillon de tissu suspect pour une analyse plus approfondie. pur determiner si une patiente à un cancer du sein a partir d'une bropsie, le tissu prélevé est examiné au microscope par un pathologiste.

I.6 Traitement

Le traitement du cancer du sein dépend du stade de la maladie, du type de cancer et de l'état de santé général de la patiente. Les principaux traitements comprennent :

- ✦ **La chirurgie :** La chirurgie est le traitement principal du cancer du sein. Elle vise à retirer la tumeur et les cellules cancéreuses environnantes. Le type de chirurgie à effectuer dépend du stade du cancer et de la localisation de la tumeur.
- ✦ **La radiothérapie :**La radiothérapie utilise des rayons X pour détruire les cellules cancéreuses restantes après la chirurgie. Elle peut être administrée avant ou après la chirurgie, ou en tant que traitement unique.
- ✦ **La chimiothérapie :**La chimiothérapie utilise des médicaments pour tuer les cellules cancéreuses. Elle peut être administrée avant ou après la chirurgie, ou en tant que traitement unique.
- ✦ **La thérapie ciblée :** La thérapie ciblée bloque les voies de signalisation des cellules cancéreuses, ce qui les empêche de se développer. Elle peut être utilisée seule ou en combinaison avec d'autres traitements.
- ✦ **L'immunothérapie :** L'immunothérapie stimule le système immunitaire pour qu'il combatte les cellules cancéreuses. Elle peut être utilisée seule ou en combinaison avec d'autres traitements.

I.7 Facteurs de risques

Les facteurs de risque du cancer du sein sont divers et peuvent être classés en plusieurs catégories, y compris les facteurs génétiques, environnementaux, sociaux et culturels. Voici une liste de certains de ces facteurs, classés par catégorie :

I.7.1 Facteurs génétiques

① **Antécédents familiaux :** La présence de cas de cancer du sein chez les proches de premier degré (mère, sœur, fille) augmente le risque.

② **Mutations génétiques :** : Les mutations dans les gènes BRCA1 et BRCA2 sont associées à un risque accru de cancer du sein.

③ **Hérédité :** Les facteurs génétiques liés à des prédispositions familiales.

I.7.2 Facteurs environnementaux

① **Age :** Le risque de cancer du sein augmente avec l'âge.

② **Exposition aux hormones :** : Une exposition prolongée aux hormones, comme la thérapie hormonale substitutive pendant la ménopause, peut accroître le risque.

③ **Alcool et tabac :** La consommation régulière d'alcool et le tabagisme sont associés à une augmentation du risque.

④ **Obésité :** Les femmes en surpoids ou obèses, en particulier après la ménopause, ont un risque accru.

I.7.3 Facteurs socio-économiques et culturels

① **Niveau éducatif :** Des études suggèrent que des niveaux d'éducation plus bas peuvent être associés à un risque plus élevé.

② **Statut économique :** Les disparités socio-économiques peuvent influencer l'accès aux soins médicaux préventifs.

③ **Stress et style de vie :** Le stress et un mode de vie stressant peuvent contribuer au risque de cancer du sein.

I.7.4 Facteurs reproductifs

① **Age des premières régles (Ménarche) :** Une ménarche précoce (avant 11-12 ans) est associée à un risque accru.

② **Age au premier accouchement :** Un premier accouchement à un âge avancé (après 30 ans) est un facteur de risque.

③ **Ménopause tardive :** Une ménopause survenant à un âge plus tardif (après 55 ans) peut augmenter le risque.

I.7.5 Facteurs liés au mode de vie

① **Activité physique :** Un manque d'activité physique peut augmenter le risque.

① **Régime alimentaire :** Certains régimes riches en graisses saturées ont été liés à un risque accru.

I.8 Stades du cancer du sein

Les différents stades du cancer du sein sont déterminés en utilisant la classification TNM (Tumeur, Ganglions lymphatiques, Métastases) ou la classification A.C.R (American College of Radiology). La signification clinique de ces stades est essentielle pour déterminer le pronostic et guider le choix du traitement. Voici une brève explication de ces deux classification :

I.8.1 Classification T.N.M

Classification internationale qui permet de se rendre compte du stade d'un cancer. La lettre T est l'initiale de tumeur et correspond à la taille de la tumeur ,la lettre N est l'initiale de node qui signifie ganglion en anglais et indique si des ganglions lymphatiques ont été ou non envahis, la lettre M est l'initiale de métastase* et signale la présence ou l'absence de métastases. [3]
Voici un aperçu de la classification TNM :

T (Tumeur) :

- Tx : La tumeur ne peut pas être évaluée.
- T0 : Pas de tumeur visible.
- T1 : La tumeur mesure moins de 2 cm de diamètre.
- T2 : La tumeur mesure entre 2 et 5 cm de diamètre.
- T3 : La tumeur mesure plus de 5 cm de diamètre.
- T4 : La tumeur s'est propagée à la paroi thoracique ou à la peau.

N (Ganglions lymphatiques) :

- Nx : L'infiltration des ganglions lymphatiques ne peut pas être évaluée.
- N0 : Pas d'envahissement des ganglions lymphatiques.
- N1 : Atteinte d'un seul ganglion lymphatique axillaire.
- N2 : Atteinte de 2 à 3 ganglions lymphatiques axillaires.
- N3 : Atteinte de 4 ou plus de ganglions lymphatiques axillaires, ou atteinte de ganglions lymphatiques mammaires internes ou sus-claviculaires.

M (Métastases) :

- M0 : Pas de métastases.
- M1 : Métastases présentes.

Combinaison des critères TNM : La combinaison des trois critères TNM permet de définir le stade du cancer du sein. Par exemple, un cancer du sein classé T2N0M0 est un cancer de stade I.
Stades du cancer du sein selon la classification TNM [6] :

- Stade 0 : Cancer in situ (non invasif).
- Stade I : Cancer invasif de petite taille, sans atteinte des ganglions lymphatiques.
- Stade II : Cancer invasif de plus grande taille, avec ou sans atteinte des ganglions lymphatiques.
- Stade III : Cancer invasif de grande taille, avec atteinte des ganglions lymphatiques ou de la peau.
- Stade IV : Cancer métastatique (propagation à d'autres organes).

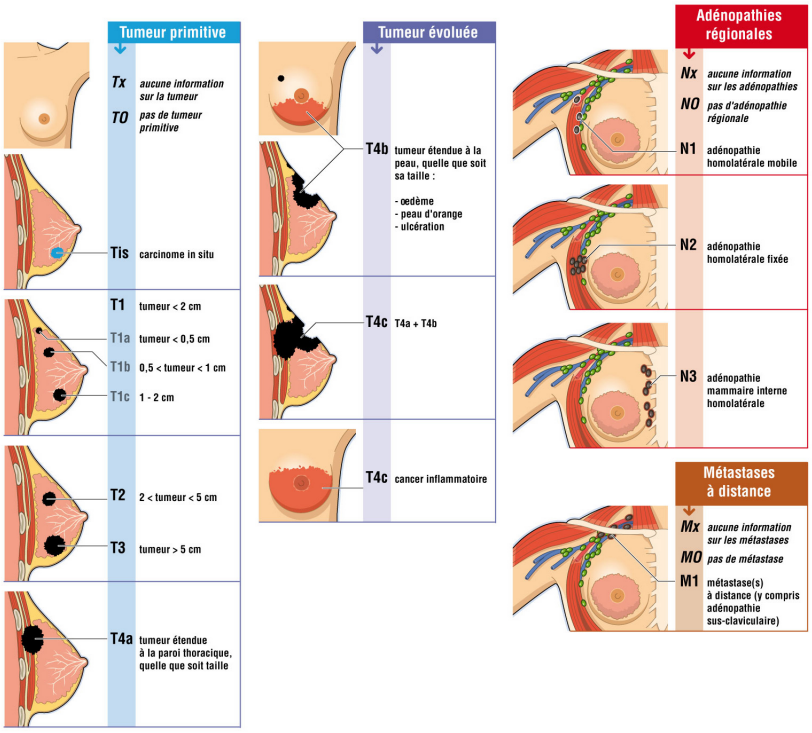


FIGURE I.4 – Classification TNM

I.8.2 Classification A.C.R

La classification Birads (Breast Imaging Reporting And Data System) de l'ACR est une classification internationale établie par l'American college of Radiology. Elle permet d'établir une attitude commune en fonction d'une anomalie dépistée en imagerie mammaire. Elle peut être unilatérale ou bilatérale en fonction de l'étude d'un ou des deux seins. [7]

- ACR0 : Classification d'attente.
- ACR1 : Mammographie normale.
- ACR2 : Présence d'anomalies bénignes sans surveillance.
- ACR3 : Anomalie bénigne avec surveillance.
- ACR4 : Présence d'une anomalie suspecte qui nécessite un prélèvement.
- ACR5 : Présence d'une anomalie évocatrice d'un cancer qui nécessite un prélèvement.

I.9 IA et cancer du sein :

L'intelligence artificielle (IA) a le potentiel de révolutionner le diagnostic et la prévention du cancer du sein. Dans le domaine du diagnostic, l'IA peut être utilisée pour analyser des images médicales, telles que des mammographies et des images histopathologiques, afin de détecter les cancers du sein de manière plus précoce et plus précise que les méthodes traditionnelles.

L'IA peut également être utilisée pour aider les médecins à prendre des décisions de traitement. Par exemple, l'IA peut être utilisée pour analyser les données génétiques d'une patiente afin de déterminer le type de cancer du sein dont elle est atteinte et le meilleur traitement à suivre.

Dans le domaine de la prévention, l'IA peut être utilisée pour développer de nouveaux tests de dépistage du cancer du sein. Par exemple, l'IA peut être utilisée pour analyser des données génétiques afin de prédire le risque de cancer du sein d'une femme.

Voici quelques exemples concrets de l'apport de l'IA au diagnostic et à la prévention du cancer du sein :

- En 2021, une étude publiée dans la revue Nature Medicine a montré qu'un système d'IA développé par Google était capable de détecter le cancer du sein avec une précision de 99,9 % sur des images de mammographie.
- En 2022, une étude publiée dans la revue Nature Communications a montré qu'un système d'IA développé par des chercheurs de l'Université de Stanford était capable de prédire le risque de cancer du sein avec une précision de 80 %.

Ces résultats sont prometteurs et suggèrent que l'IA a le potentiel de sauver des vies en aidant à détecter et à prévenir le cancer du sein plus tôt.

Cependant, il est important de noter que l'IA n'est pas encore prête à remplacer les médecins. L'IA est une technologie puissante qui peut être utilisée pour aider les médecins à prendre des décisions plus éclairées, mais elle ne peut pas remplacer leur expertise et leur jugement.

Dans un avenir proche, l'IA est susceptible de jouer un rôle de plus en plus important dans le diagnostic et la prévention du cancer du sein. L'IA a le potentiel de rendre le dépistage du cancer du sein plus précoce et plus précis, et de permettre aux médecins de prendre des décisions de traitement plus personnalisées.

I.10 Conclusion

En conclusion, le cancer du sein reste un défi de santé publique majeur, nécessitant une compréhension approfondie de ses multiples facettes. La sensibilisation aux facteurs de risque, aux symptômes et aux méthodes de dépistage reste essentielle pour promouvoir la détection précoce.

Cependant, à l'ère de l'innovation technologique, nous pouvons envisager des avancées significatives dans le domaine du diagnostic du cancer du sein. Les progrès rapides dans les technologies, notamment le machine learning et l'intelligence artificielle, ouvrent de nouvelles perspectives pour une détection plus précise et précoce.

L'intégration du machine learning dans l'analyse d'imagerie médicale, telle que la mammographie, permet une identification plus rapide des anomalies, réduisant ainsi le temps nécessaire pour poser un diagnostic. Ces technologies offrent également la possibilité de personnaliser les diagnostics en fonction des caractéristiques individuelles,améliorant ainsi l'efficacité des traitements.

En investissant dans la recherche et le développement de solutions basées sur le machine learning, nous pouvons espérer une révolution dans la manière dont nous abordons le cancer du sein. Cela pourrait non seulement conduire à des diagnostics plus précoces et plus précis, mais également à des interventions thérapeutiques plus ciblées et personnalisées.

Ainsi, tandis que nous continuons à promouvoir la sensibilisation et les méthodes traditionnelles de dépistage, nous devons également embrasser les opportunités que la technologie moderne offre. Le mariage entre la science médicale et les avancées technologiques peut être la clé pour réduire l'impact du cancer du sein et améliorer la qualité de vie des personnes touchées. En unissant nos efforts, nous pouvons aspirer à un avenir où le diagnostic et le traitement du cancer du sein sont plus efficaces et accessibles.

ChapitreII

Apprentissage automatique et apprentissage profond

II.1 Introduction

Au cours des dernières décennies, le domaine de la technologie a connu une évolution remarquable, introduisant des concepts tels que l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond. Ces avancées ont profondément transformé notre capacité à traiter les données complexes, ouvrant de nouvelles perspectives dans divers domaines, notamment en médecine, où elles sont appliquées avec succès pour aider le diagnostic et la prédiction du cancer du sein.

L'intelligence artificielle représente un domaine vaste et dynamique où les ordinateurs sont formés pour manifester un comportement intelligent à travers une variété de techniques statistiques et d'optimisation. Dans ce contexte, l'apprentissage automatique est une sous-discipline de l'intelligence artificielle, se distingue par sa capacité à permettre aux machines d'apprendre à partir de données sans être explicitement programmées. Cependant, c'est l'évolution plus récente de l'apprentissage profond qui a véritablement révolutionné notre approche de l'analyse des données médicales.

Dans ce chapitre, nous explorerons les différents types d'algorithmes d'apprentissage automatique et d'apprentissage profond utilisés pour le diagnostic et la prédiction du cancer du sein.Nous analyserons les avantages et les limitations de chaque approche, ainsi que leurs applications pratiques dans le domaine médical. En mettant l'accent sur les récents développements et les avancées significatives, nous illustrerons l'importance croissante de l'apprentissage automatique et de l'apprentissage profond dans la lutte contre le cancer du sein, et leur potentiel à révolutionner la manière dont cette maladie est diagnostiquée, traitée et gérée.

II.2 Apprentissage automatique

II.2.1 Définition de l'apprentissage automatique

Le ML est une branche de l'intelligence artificielle qui permet aux systèmes informatiques d'apprendre et de s'améliorer à partir de données, sans être explicitement programmés. Il utilise des algorithmes et des modèles statistiques pour analyser des motifs dans les données et faire des prédictions ou des décisions.[8]

II.2.2 Les raisons d'utiliser l'apprentissage automatique

L'apprentissage automatique présente de nombreux avantages et applications dans divers domaines.
Voici quelques raisons principales pour lesquelles il est largement utilisé :

① **Capacité à traiter de grandes quantités de données :** L'apprentissage automatique est particulièrement efficace pour analyser et interpréter de vastes ensembles de données qui dépassent les capacités de traitement humain.

② **Identification de modèles complexes :** Il peut découvrir des motifs, des tendances et des relations non triviales dans les données qui peuvent échapper à l'observation humaine ou à l'analyse statistique traditionnelle.

③ **Détecter des anomalies :** Le ML peut être utilisé pour détecter des anomalies dans les données, ce qui peut aider à prévenir les fraudes et les erreurs.

④ **Amélioration de la prise de décision :** Le ML peut aider à de meilleurs prise de décisions en fournissant des informations précises et prédictives. par exemple : Segmenter les clients, Minimiser les couts, Optimiser les profits, Diagnostiquer des maladies.

Donc l'apprentissage automatique offre beaucoup d'avantages significatifs qui en font une technologie précieuse dans de nombreux domaines, de l'informatique à la médecine en passant par le commerce et bien d'autres encore.

II.2.3 Types de systémes d'apprentissage automatique

Il existe plusieurs types de ML classé en grande catégorie [8] :

- En fonction de la présence ou de l'absence de supervision humaine (apprentissage supervisé, non supervisé, semi-supervisé ou avec renforcement).
- En fonction de savoir si l'apprentissage se produit de manière continue au fil du temps ou non (apprentissage en ligne ou apprentissage groupé).
- En fonction de comparer les nouvelles données à des données connues, ou qu'il détecte au contraire des éléments de structuration dans les données d'entraînement et construise un modèle prédictif à la façon d'un scientifique (apprentissage à partir d'observations ou apprentissage à partir d'un modèle).

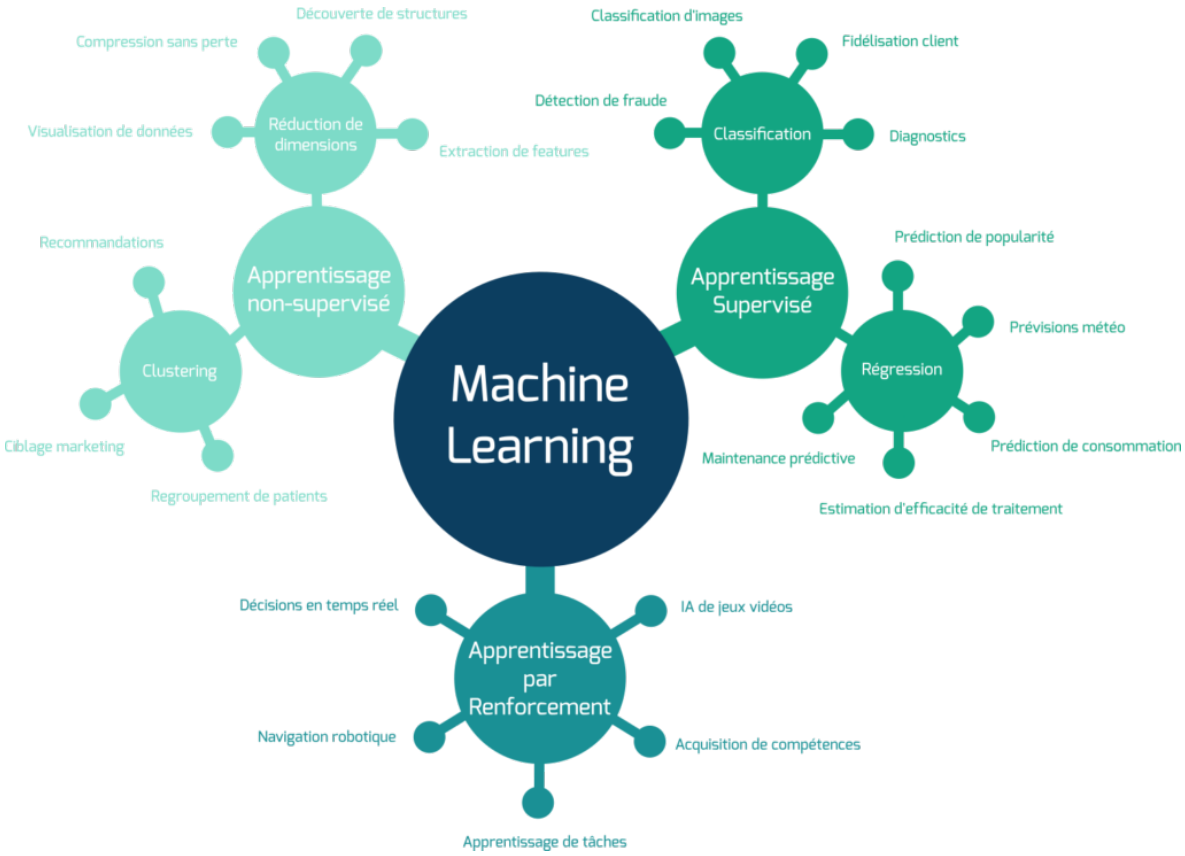


FIGURE II.1 – Les types d'apprentissage automatique

II.2.3.1 Apprentissage supervisé

L'apprentissage supervisé, également appelé apprentissage automatique supervisé, est une sous-catégorie de l'apprentissage automatique et de l'intelligence artificielle. Il se caractérise par l'utilisation de jeux de données étiquetés qui entraînent des algorithmes permettant de classer des données ou de prédire des résultats avec précision. Au fur et à mesure que les données en entrée sont introduites dans le modèle, celui-ci ajuste ses pondérations jusqu'à ce que le modèle soit correctement ajusté. C'est le processus de validation croisée. Avec l'apprentissage supervisé, les organisations peuvent résoudre divers problèmes du monde réel à grande échelle, comme la classification des courriers indésirables dans un dossier distinct de votre boîte de réception. [9]

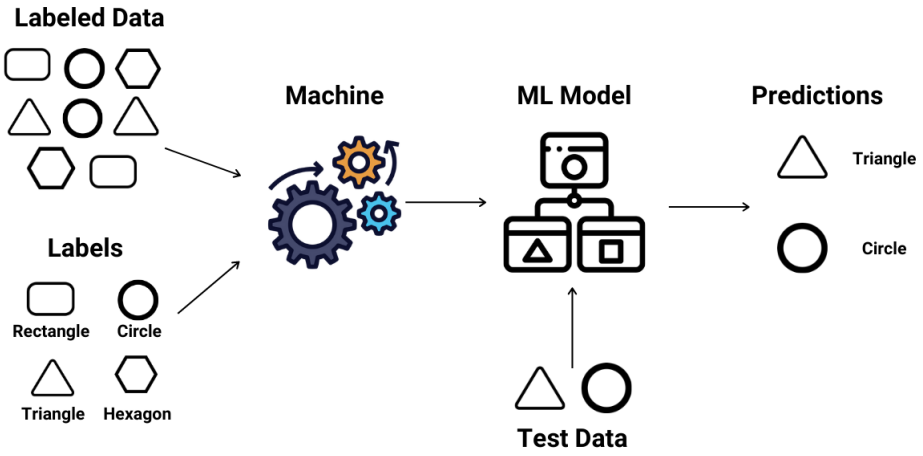


FIGURE II.2 – Apprentissage supervisé

II.2.3.1.1 Les modèles d'apprentissage supervisé Plusieurs modèles que l'on peut implémenter sous forme d'algorithmes (mathématiques puis informatiques) existent en apprentissage supervisé et diffèrent par leur manière d'aborder l'entraînement sur les données mais aussi le type de label à prédire (valeur continue, classe...).

- **La régression :** Est l'une des techniques d'apprentissage supervisé les plus populaires dans la prédiction d'une valeur continue. Par exemple, on peut utiliser ce modèle pour prédire le prix d'une maison sachant sa taille, le nombre de chambres et le lieu où elle se trouve.
- **La classification :** Dans les tâches de classification, le programme d'apprentissage automatique doit tirer une conclusion à partir des valeurs observées et déterminer à quelle catégorie appartiennent les nouvelles observations. Par exemple, lors du filtrage des e-mails comme "spam" ou "non spam", le programme doit examiner les données d'observation existantes et filtrer les e-mails en conséquence. [10]

II.2.3.2 Apprentissage non supervisé

L'apprentissage non supervisé, utilise des algorithmes d'apprentissage automatique pour analyser et regrouper des jeux de données non étiquetés. Ces algorithmes découvrent des motifs cachés ou des groupements de données sans nécessiter d'intervention humaine. Sa capacité à découvrir les similitudes et les différences d'informations en fait la solution idéale pour l'analyse d'exploration des données, les stratégies de vente croisée, la segmentation de la clientèle et la reconnaissance d'images.[9]

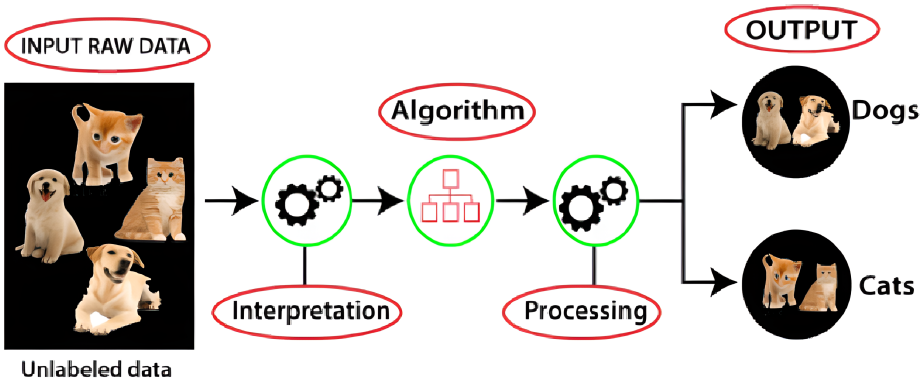


FIGURE II.3 – Apprentissage non supervisé

II.2.3.2.1 Les modèles d'apprentissage non supervisé Les modèles d'apprentissage non supervisé sont utilisés pour trois tâches principales : le regroupement(Clustering), l'association et la réduction de la dimensionnalité. Nous définirons ci-dessous chaque méthode d'apprentissage et mettrons en évidence les algorithmes communs et les approches permettant de les mener efficacement.

- **Clustering :** Les données non étiquetées sont regroupées à l'aide de techniques de regroupement en fonction de leurs similitudes ou de leurs différences. Par exemple, si une équipe travaille sur la segmentation du marché, l'algorithme de clustering k-moyennes attribuera des points de données similaires aux groupes qui représentent un ensemble de paramètres. Le regroupement peut se faire en fonction de l'emplacement, des niveaux de revenus, de l'âge des acheteurs ou de n'importe quelle autre variable.
- **Association :** La méthode d'association de l'apprentissage non supervisé est intéressante pour trouver des relations entre les variables d'un jeu de données. C'est la technique utilisée pour créer le message de type « les autres clients ont également consulté ». Elle est particulièrement adaptée aux moteurs de recommandation. Si 15 clients ayant acheté un nouveau téléphone ont également commandé un casque, les algorithmes recommandent un casque à tous les clients qui mettent un téléphone dans leur panier.
- **Réduction de la dimensionnalité :** Il arrive qu'un jeu de données comporte un nombre de caractéristiques exceptionnellement élevé. La réduction de la dimensionnalité permet de réduire ce nombre sans compromettre l'intégrité des données. Il s'agit d'une technique couramment utilisée avant le traitement des données. Cela sert par exemple à supprimer le bruit d'une image pour améliorer sa qualité. [11]

II.2.3.3 Apprentissage semi supervisé

L'apprentissage semi-supervisé est une technique d'apprentissage automatique qui consiste à partir d'un jeu de données constitué en majorité de données non labellisées et en minorité de données labellisées. Il se situe entre l'apprentissage supervisé qui utilise des données labellisées et l'apprentissage non supervisé qui utilise des données non labellisées.[10]

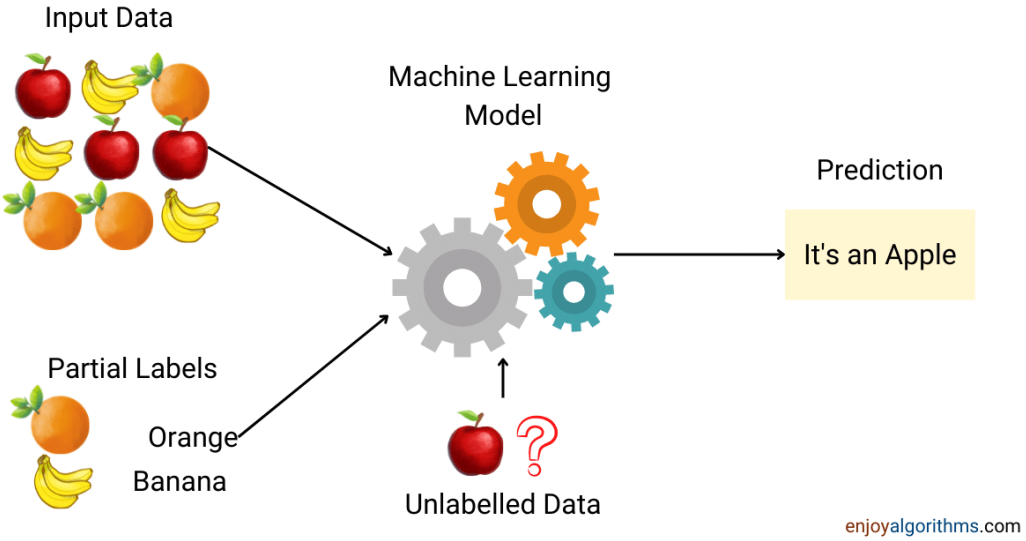


FIGURE II.4 – Apprentissage semi supervisé

II.2.3.3.1 Modèles d'apprentissage semi-supervisé Il existe de nos jours plusieurs méthodes qui ont été développées par des grands chercheurs pour résoudre des problèmes d'apprentissage semi-supervisé notamment les méthodes de régularisation de consistance parmi lesquelles l'algorithme FixMatch développé par Sohn et al. et de l'algorithme The Mean Teacher développé par Antti Tarvainen et Harri Valpola de The Curious AI Company que nous allons présenter par la suite. Ces méthodes ont obtenu des performances remarquables sur des problèmes de classification avec des jeux de données comme CIFAR-10 par rapport à ce qui était connu auparavant dans l'état de l'art.[10]

II.2.3.4 Apprentissage par renforcement

L'apprentissage par renforcement est un système où un agent observe l'environnement, sélectionne et accomplit des actions, et reçoit des récompenses en retour. L'agent apprend à identifier la meilleure stratégie pour maximiser les récompenses, définie par une politique. Les robots utilisent souvent des algorithmes d'apprentissage par renforcement pour apprendre à marcher. Un exemple célèbre est le programme AlphaGo de DeepMind, qui a battu le champion de go Ke Jie en mai 2017 en s'entraînant sur des millions de parties et en jouant contre lui-même. Pendant les parties contre Ke Jie, AlphaGo appliquait uniquement la politique apprise.[8]

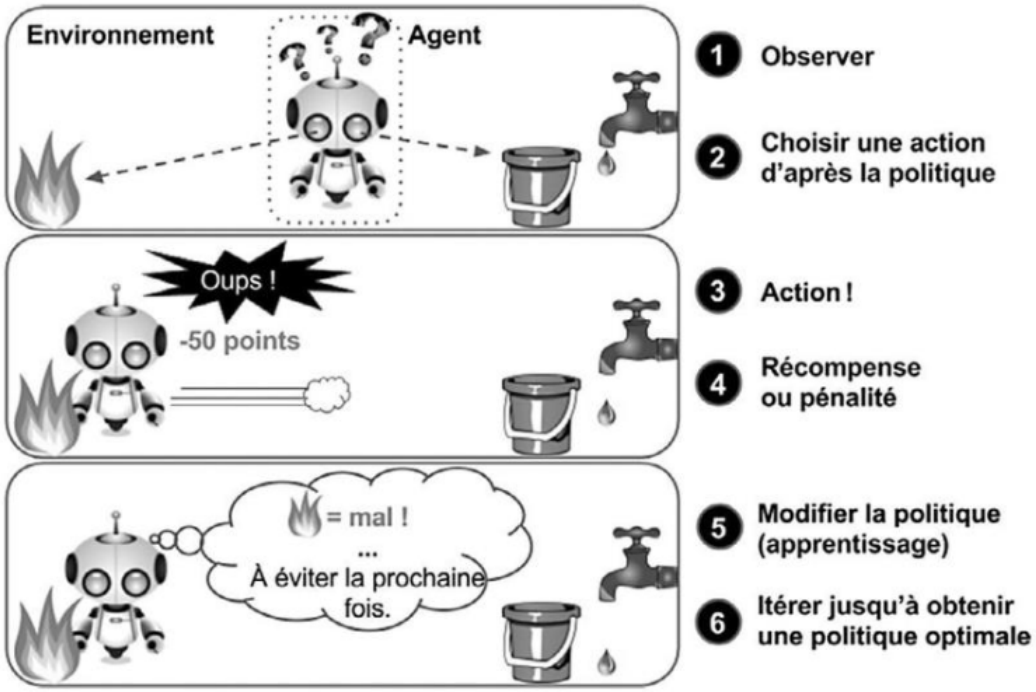


FIGURE II.5 – Apprentissage par renforcement

II.2.3.5 La différences entre les types d'apprentissage automatique

Voici un tableau comparatif des différents types d'apprentissage en machine :

Caractéristique	**Apprentissage supervisé**	**Apprentissage non supervisé**	**Apprentissage semi-supervisé**	**Apprentissage par renforcement**
Type de données d'entrée	Données étiquetées	Données non étiquetées	Mélange de données étiquetées et non étiquetées	Données d'entrée et récompenses/retours
Objectif principal	Prédire une sortie	Identifier des structures	Utiliser à la fois des données étiquetées et non étiquetées pour prédire une sortie	Maximiser une récompense cumulative
Exemples d'applications	Classification, Régression	Clustering, Réduction de dimensionnalité	Classification semi-supervisée, Régression semi-supervisée	Apprentissage autonome d'un agent à partir d'expériences
Méthodes courantes	Régression linéaire, Arbres de décision	K-Means, Analyse en composantes principales (PCA)	Propagation de labels, Autoencodeurs semi-supervisés	Q-Learning, Deep Q-Networks (DQN)
Évaluation	Précision, Erreur	Cohérence, Silhouette	Métriques combinant à la fois les données étiquetées et non étiquetées	Score de récompense, Valeur de la politique
Avantages	Modèles précis avec des données étiquetées	Exploration des structures cachées dans les données	Utilise efficacement les données non étiquetées	Capacité à apprendre par l'interaction avec l'environnement
Inconvénients	Dépendance aux données étiquetées de haute qualité	Dépendance à la qualité et à la représentativité des données non étiquetées	Nécessite un bon équilibre entre données étiquetées et non étiquetées	Sujette à des problèmes de stabilité et de convergence

TABLE II.1 – Comparaison des types d'apprentissage en machine

II.2.4 Les principaux difficultés de l'apprentissage automatique

La principale tache de ML est de sélectionner un algorithme et l'entrainer sur certaines données, cela implique parfois un mauvais choix de l'algorithme et parfois de données inappropriées.
Voici quelque exemples de difficultés :

II.2.4.1 Données d'apprentissage en nombre insuffisant

La plupart des algorithme d'apprentissage automatique requièrent une grande quantité de données pour fonctionner correctement, même pour de simple problème, il faut un grand nombre de données pour résoudre des problèmes complexes tels que la reconnaissance vocale et le diagnostic médicale.

II.2.4.2 Données d'entraînement non représentatives

Il est crucial d'avoir un ensemble de données d'entraînement représentatif pour généraliser efficacement un modèle. Par exemple, un modèle linéaire entraîné sur des données de bonheur et de richesse s'améliore lorsqu'on ajoute des pays manquants, mais il reste imparfait car la relation entre richesse et bonheur n'est pas linéaire pour les pays très riches ou très pauvres. Obtenir un ensemble de données d'entraînement représentatif est souvent difficile à cause du bruit et du biais d'échantillonnage.[8]

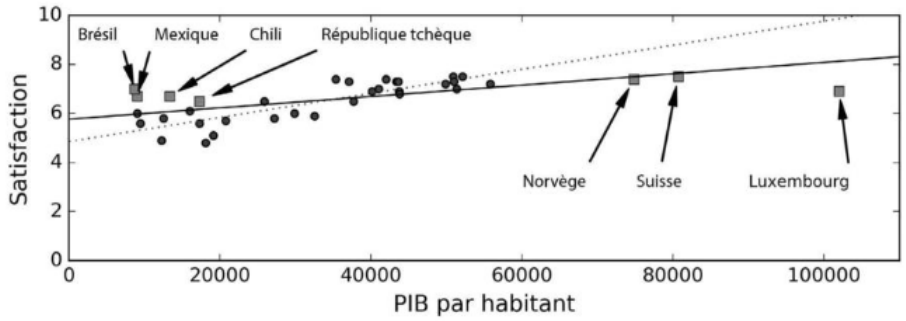


FIGURE II.6 – Un échantillon d'entraînement plus représentatif

II.2.4.3 Données de mauvaise qualité

Si le jeu d'entrainement contient trop d'erreur, de données aberrante et de bruit (dû à la mauvaise qualité des mesures), le modèle aura du mal a détecter les structures cachée donc il aura moins de chance a retourner de bons résultats. Pour remédier a cela il faut consacrer du temps à nettoyer vos données d'apprentissage.
Les exemples suivants illustrent les cas dans lesquels il faudra nettoyer les données d'entraînement :

① Si certaines données sont manifestement aberrantes, il peut être utile de les supprimer ou d'essayer de corriger manuellement les erreurs.

② S'il manque quelques valeurs dans certaines observations , vous devez décider si vous voulez ignorer totalement cette variable, ou ignorer ces observations, ou remplir ces

valeurs manquantes (par exemple par la moyenne), ou entraîner un modèle avec cette variable et un autre sans.

II.2.4.4 Variables non pertinentes

Si les données d'entrainement sont mauvaises, la sortie le sera aussi. Le modèle n'aura pas la capacité d'apprendre sauf si les données d'entrés contiennent suffisamment de variables pertinentes et pas trop de non pertinentes. Le succès d'un projet de Machine Learning dépend largement du choix judicieux des variables d'entraînement, un processus connu sous le nom d'ingénierie des variables (feature engineering en anglais). Cela implique les étapes suivantes :

① La sélection de variables, processus consistant à sélectionner parmi les variables existantes celles sur lesquelles il est le plus utile de réaliser l'entraînement.

② L'extraction de variables, combinant plusieurs variables existantes pour en produire une autre plus utile.

③ Introduction de nouvelles variables grâce à la collecte de nouvelles données.

II.2.4.5 Surajustement des données d'entraînement

Le surajustement (overfitting) en machine learning survient lorsqu'un modèle fonctionne bien sur les données d'apprentissage, mais pas sur de nouvelles données, en raison d'une généralisation excessive. Cela peut arriver lorsque les modèles complexes détectent des structurations dans le bruit des données, surtout si le jeu d'entraînement est petit ou bruyant. Pour éviter le surajustement, il est conseillé de simplifier le modèle en réduisant le nombre de paramètres ou d'attributs, de rassembler plus de données d'apprentissage, et de réduire le bruit dans les données en corrigeant les erreurs et en supprimant les valeurs aberrantes.[8]

II.2.4.6 Sous-ajustement des données d'entraînement

Comme vous pouvez vous en douter, le sous-ajustement (en anglais, underfitting) est l'opposé du surajustement : il se produit lorsque votre modèle est trop simple pour découvrir la structure sous-jacente des données. Ainsi, un modèle linéaire de satisfaction individuelle risque de sous-ajuster : la réalité est tout simplement plus complexe que le modèle, c'est pourquoi ses prédictions risquent d'être inexactes, même sur les exemples d'apprentissage.[8]
Voici quelque solution pour éviter le underfitting :

① Choisir un modèle plus puissant, avec plus de paramètres.

② Fournir de meilleures variables à l'algorithme d'apprentissage (en les transformant au besoin).

③ Réduire les contraintes sur le modèle (p. ex. en réduisant l'hyperparamètre de régularisation).

II.3 Les principaux algorithmiques de l'apprentissage automatique

Voici quelques algorithmique de l'apprentissage supervisé :

II.3.1 Régression linéaire/logistique

La régression linéaire et la régression logistique sont deux algorithmes couramment utilisés pour modéliser la relation entre une variable de sortie et une ou plusieurs variables d'entrée. Bien qu'elles partagent des similitudes, il existe des différences fondamentales entre les deux approches, notamment en ce qui concerne le type de variable de sortie qu'elles peuvent gérer et les types de prédictions qu'elles génèrent.

Caractéristique	**Régression linéaire**	**Régression logistique**
Type de variable de sortie	Continue	Catégorielle
Type de prédiction	Valeur numérique	Probabilité
Fonctionnement interne	Équation linéaire	Fonction logistique
Exemples d'applications	Prédiction du prix d'une maison, estimation de la température, modélisation du nombre de buts marqués	Classification des emails comme spam ou non, prédiction du risque de crédit, détection de fraude

TABLE II.2 – Comparaison entre la régression linéaire et logistique

II.3.2 Machines à vecteurs de support (SVM)

Les SVMs sont une famille d'algorithmes d'apprentissage automatique qui permettent de résoudre des problèmes tant de classification que de régression ou de détection d'anomalie. Ils sont connus pour leurs solides garanties théoriques, leur grande flexibilité ainsi que leur simplicité d'utilisation même sans grande connaissance de data mining.
Les SVMs ont été développés dans les années 1990. Comme le montre la figure ci-dessous, leur principe est simple : il ont pour but de séparer les données en classes à l'aide d'une frontière aussi « simple » que possible, de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale. Cette distance est aussi appelée « marge » et les SVMs sont ainsi qualifiés de « séparateurs à vaste marge », les « vecteurs de support » étant les données les plus proches de la frontière.[18]

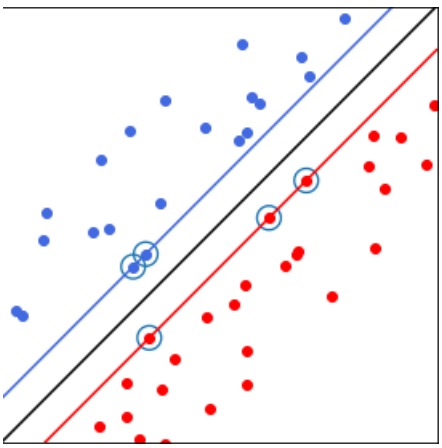


FIGURE II.7 – Machines à vecteurs de support

II.3.3 Forêts aléatoires

Forêt aléatoire est un algorithme d'apprentissage automatique couramment utilisé, et breveté par Leo Breiman et Adele Cutler, qui permet d'assembler les sorties de plusieurs arbres de décision pour atteindre un résultat unique. Sa souplesse d'utilisation et sa flexibilité ont favorisé son adoption, car il gère à la fois les problèmes de classification et de régression.[19]
L'algorithme de forêt aléatoire présente un nombre d'avantages et de défis clés lorsqu'il est utilisé pour résoudre les problèmes de classification ou de régression.
En voici quelques-uns :

- ✦ Réduction du risque de surajustement : Les arbres de décision peuvent produire un surajustement, car ils ont tendance à ajuster tous les échantillons à partir des données d'entraînement.
- ✦ Flexibilité : Comme la forêt aléatoire peut gérer à la fois des tâches de régression et de classification avec un haut degré d'exactitude, il s'agit d'une méthode appréciée des spécialistes des données.
- ✦ Détermination facile de l'importance de la fonction : Une forêt aléatoire permet d'évaluer facilement l'importance ou la contribution des variables au modèle.[19]

II.3.4 K plus proche voisin(KNN)

KNN est un algorithme qui reçoit un ensemble de données qui est étiqueté avec des valeurs de sorties correspondantes sur lequel il va pouvoir s'entraîner et définir un modèle de prédiction. Cet algorithme pourra par la suite être utilisé sur de nouvelles données afin de prédire leurs valeurs de sorties correspondantes.[20]

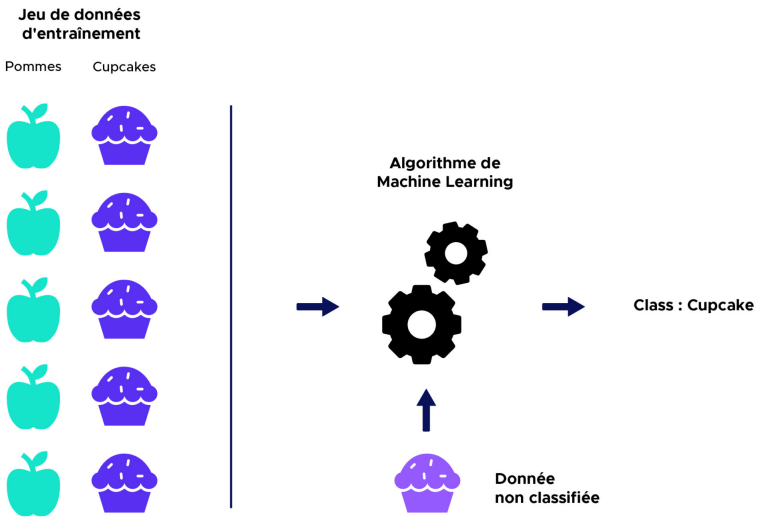


FIGURE II.8 – KNN

Quelques algorithme de l'apprentissage non supervisé :

II.3.5 K-means

L'algorithme K-means est l'algorithme de classification le plus utilisé qui utilise une mesure de distance explicite pour partitionner l'ensemble de données en clusters.
Le concept principal de l'algorithme K-mean est de représenter chaque cluster par le vecteur des valeurs d'attribut moyennes de toutes les instances d'apprentissage pour les attributs numériques et par le vecteur des valeurs modales (les plus fréquentes) pour les attributs nominaux qui sont affectés à ce cluster. Cette représentation de cluster est appelée centre de cluster.[21]

II.3.6 Algorithme Apriori

L'algorithme Apriori est un algorithme d'exploration de données conçu en 1994, par Rakesh Agrawal et Ramakrishnan Sikrant, dans le domaine de l'apprentissage des règles d'association. Il sert à reconnaitre des propriétés qui reviennent fréquemment dans un ensemble de données et d'en déduire une catégorisation.[22]

L'algorithme Apriori s'exécute en deux étapes[22] :

1) Génération de tous les itemsets fréquents, c'est-à-dire

$$IF = \{X_i \subseteq T \,|\, \text{supp}(X_i) = X_i.\text{count} \geq \text{minsupp}, i = 1, 2, ..n\}$$

2) Génération de toutes les règles d'associations de confiance à partir des itemsets fréquents, c'est-à-dire

$$\{X_i, Y_j \subseteq IF \,|\, X_i \cap Y_j = \varnothing \wedge \text{Conf}(X_i \rightarrow Y_j) \geq \text{minconf}\, i = 1, 2, ..p\, j = 1, 2, ..q\}$$

II.4 Apprentissage profond

L'apprentissage profond est un sous-domaine de l'apprentissage automatique traitant des algorithmes inspirés par la structure du cerveau humain. En d'autres termes, il reflète la façon dont notre cerveau fonctionne. Les algorithmes de l'apprentissage profond sont similaires à la façon dont le système nerveux se structure où chaque neurone se connecte et transmet des informations. L'apprentissage profond est un progrès relativement nouveau dans la programmation de réseaux neuronaux et représente une façon de former des réseaux neuronaux profonds essentiellement, tout réseau neuronal avec plus de deux couches est profonde,si on entraine le modèle linéaire sur ces données, on obtient une ligne continue, tandis que l'ancien modèle est représenté par la ligne en pointillé.

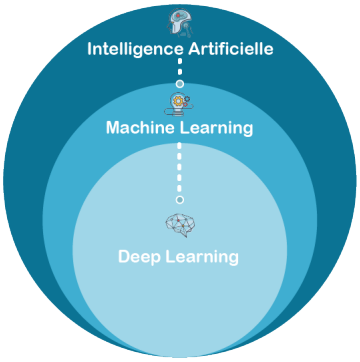


FIGURE II.9 – Apprentissage profond

II.4.1 Histoire de l'apprentissage profond

L'apprentissage automatique, dont le terme a été popularisé en 1959 par Arthur Samuel, a connu une longue évolution avant de déboucher sur les réseaux de neurones profonds d'aujourd'hui'hui. Partant des premières représentations hiérarchiques multicouches d'Ivakhnenko dans les années 60, l'avancée décisive fut l'implémentation de la rétro-propagation par Hinton, Rumelhart et Williams en 1986, permettant d'entraîner efficacement les réseaux profonds. Malgré un ralentissement dans les années 60-70, les progrès conceptuels et l'accroissement de la puissance de calcul ont finalement mené à la percée d'AlexNet en 2012. Ce réseau révolutionnaire d'apprentissage profond créé par Krizhevsky et Sutskever sous la supervision de Hinton a démontré des performances inégalées en reconnaissance d'images, relançant l'essor actuel de l'intelligence artificielle.[40]

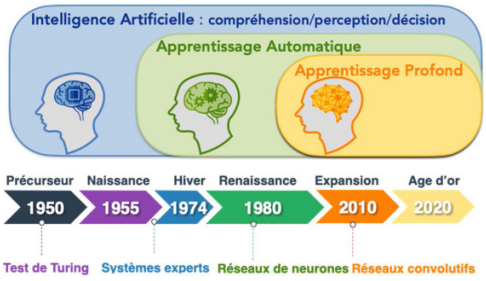


FIGURE II.10 – L'évolution de l'apprentissage profond

II.4.2 Domaine d'application de l'apprentissage profond

Le deep Learning est utilisé dans de nombreux domaines :

- ✦ Reconnaissance d'image.
- ✦ Traduction automatique.
- ✦ Voiture autonome.
- ✦ Diagnostic médical.
- ✦ Recommandations personnalisées.
- ✦ Modération automatique des réseaux sociaux.
- ✦ Prédiction financière et trading automatisé.
- ✦ Identification de pièces défectueuses.
- ✦ Détection de malwares ou de fraudes.
- ✦ Chatbots (agents conversationnels).
- ✦ Robots intelligents.

C'est aussi grâce au deep Learning que l'intelligence artificielle de Google Alpha Go a réussi à battre les meilleurs champions de Go en 2016. Le moteur de recherche du géant américain est lui-même de plus en plus basé sur l'apprentissage par deep Learning plutôt que sur des règles écrites. [13]

II.4.3 Réseaux de neurones (ANN)

II.4.3.1 Neurone biologique

Les neurones sont des cellules cérébrales uniques, composées d'un corps cellulaire contenant le noyau, de courts prolongements appelés dendrites et d'un long axone. L'axone se ramifie en télodendrons terminés par des synapses, qui se connectent à d'autres neurones. La communication neuronale s'effectue par des impulsions électriques (potentiels d'action) le long de l'axone et des signaux chimiques (neurotransmetteurs) libérés aux synapses. Un neurone déclenche ses propres impulsions lorsqu'il reçoit suffisamment de neurotransmetteurs. Bien que chaque neurone soit relativement simple, leur organisation en vastes réseaux permet des calculs extrêmement complexes, comparable à une fourmilière élaborée construite par de simples fourmis.[12]

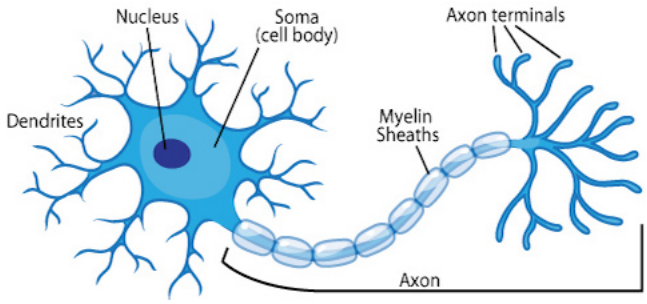


FIGURE II.11 – Neurone Biologique

II.4.3.2 Perceptron

Inventé en 1957 par Frank Rosenblatt au laboratoire aéronautique de Cornell, le perceptron est un réseau de neurones artificiels basé sur les premiers concepts de neurones artificiels. Il permet à l'ordinateur d'apprendre grâce à de nouvelles données en réalisant des calculs pour détecter des modèles dans les données d'entrée. Considéré comme le réseau de neurones artificiels le plus simple, il est utilisé dans le Deep Learning pour la formation supervisée de classificateurs binaires. Cet algorithme joue un rôle crucial dans les projets de Machine Learning en organisant et en classifiant les données.[14]

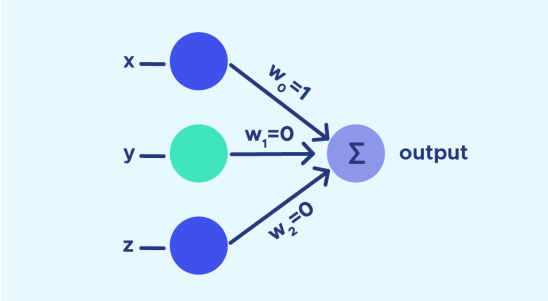


FIGURE II.12 – Perceptron

II.4.3.3 Neurone Formel (Artificiel)

Le terme "neurone formel" ou "neurone artificiel" désigne une unité de base dans les réseaux de neurones artificiels, simulant les fonctions des neurones biologiques. Un neurone formel prend plusieurs entrées pondérées, les somme, puis applique une fonction d'activation (comme sigmoïde, ReLU, ou tangente hyperbolique) pour produire une sortie.
Mathématiquement, la sortie d'un neurone formel peut être représentée comme suit :
$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$

Où :

- y est la sortie du neurone.
- f est la fonction d'activation.
- x_i sont les entrées du neurone.
- w_i sont les poids associés à chaque entrée.

✦ b est le biais.

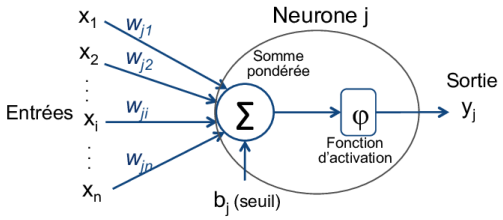


FIGURE II.13 – Structure d'un neurone artificiel

II.4.3.4 Fonction d'activation

Le choix de la fonction d'activation est conditionné par la nature du modèle que l'on souhaite reproduire, étant donné la diversité des fonctions d'activation disponibles.

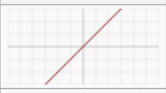
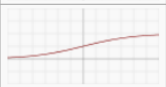
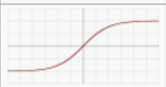
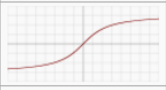
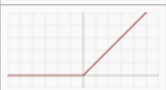
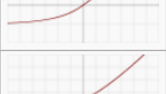
Name	Plot	Equation	Derivative
Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2+1}$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU) [2]		$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Exponential Linear Unit (ELU) [3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1 + e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$

FIGURE II.14 – Fonction d'activation

La fonction d'activation à utiliser sont généralement non-linéaire comme fonction relu ou Tangente Hyperbolique, le but de ses fonctions c'est de faire varier le résultat entre un interval précis. ex : pour Sigmoïde le résultat varie entre 0 et 1.[15]

II.4.3.5 Topologies des réseaux de neurones

Un réseau de neurones n'est rien d'autre que l'agention de plusieurs neurones artificiels ensemble, c'est alors qu'on peut parler d'architecture ou famille d'architectures. Cette dernière peut varier selon le type de problème qu'on souhaite traiter.

II.4.3.5.1 Famille des réseaux à propagation avant (Feedforward) C'est lorsque le réseau propage l'information reçut de la couche entrées vers la couche sortie sans retour en arrière.

- **Perceptron simple :** C'est un réseau qui a une seule couche qui ne comporte pas de boucle, et dont la dynamique est déclenché par la réception d'une entrée, il est dit simple car il se compose d'une couche entrée, et d'une autre sortie, l'ensemble des noeux de deux couches est connectés les uns autres.
- **Perceptron multi couche :** C'est un réseau qui contient une couche d'entrée, une couche de sortie, et une ou plusieurs couches cachées au milieu. Grâce à ces couches cachées, le réseau peut approximer n'importe quel type de résultat non-linéaire. Pour ce faire, il suffit d'ajouter suffisamment de neurones dans les couches cachées. Cette capacité d'approximation universelle fait de ce réseau un outil puissant pour modéliser des problèmes complexes.
- **Deep learning pour les réseaux profond :** Ce type peut être considéré comme un MLP avec plusieurs couche cachées qu'on retrouve dans des domaines tel que :traduction, reconnaissance du langage, traitement d'image.

II.4.3.5.2 Famille des réseaux de neurones récurrents Les réseaux de neurones récurrents (RNN) sont ainsi appelés car ils comportent des cycles dans leurs connexions, ce qui induit une dynamique particulière permettant au réseau de s'auto-entretenir et de maintenir un état interne changeant au fil du temps.

- **Modèle Hopfield :** C'est un réseau constitué d'une couche unique comprenant à la fois les entrées et les sorties, où chaque unité qui le compose est interconnectée avec toutes les autres unités.
- **Réseau de neurones récurrent à couche :** C'est un type spécifique de réseaux récurrents, où le signal d'entrée se propage vers l'avant entre les couches, mais intègre également des informations provenant des étapes précédentes.

II.4.3.5.3 Famille des réseaux à résonance

- **Machine de Boltzmann :** C'est un réseau de neurones ou toutes les couches cachées sont interconnectées, cette machine apprend un comportement désirer, c'est une extension probabiliste du modèle Hopfield.
- **Réseaux auto-organisés :** Ce type de réseau utilise une méthode d'apprentissage non supervisées, ce type de réseaux se distingue par une connectivité locale, adapté pour le traitement de l'information en spirales, le modèle le plus connue la carte auto-organisatrices de Kohonen.

II.4.3.6 Les couches d'un réseau de neurone

Le perceptron est organisé en trois couches :

- **Couche entrée(Input Layer) :** C'est l'ensemble des neurones qui reçoit le signal d'entrée du réseau, et chaque neurone de cette couche est ensuite connecté à la couche suivante.
- **Couche cachée(Hidden layers) :** Elles peuvent être une ou plusieurs, ces couches mettent en évidence les relations entre les variables. Le choix du nombre de couches et de neurones est intuitif et requiert l'expérience d'un expert.
- **Couche sortie(Output Layer) :** Elle représente le résultat du réseau de neurones c'est ce qu'on appel la prédiction.

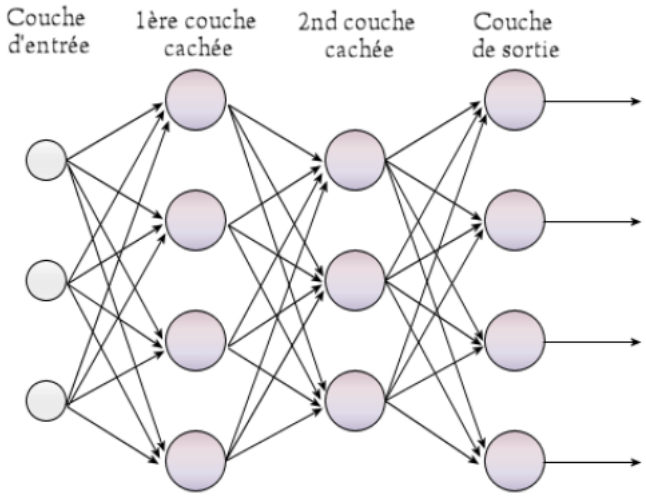


FIGURE II.15 – L'architecture d'un perceptron

II.4.4 Réseaux de neurones convolutifs

Un réseau neuronal convolutif (CNN ou ConvNet) est une architecture réseau pour le Deep Learning qui apprend directement à partir des données.Ce type de réseau est particulièrement efficace lorsqu'il s'agit de trouver des patterns dans des images afin de reconnaître des objets, des classes et des catégories. Les CNN peuvent aussi être efficaces pour la classification de données audio, de séries temporelles et de signaux.[16]

II.4.4.1 Fonctionnement du CNN

Un réseau neuronal convolutif peut disposer de dizaines, voire de centaines de couches qui apprennent chacune à détecter différentes caractéristiques d'une image. Des filtres sont appliqués à chaque image du jeu d'apprentissage avec différentes résolutions, puis la sortie de chaque image convoluée est utilisée comme entrée de la couche suivante. Au début, ces filtres peuvent concerner des caractéristiques très simples, comme la luminosité et les contours, puis se complexifier jusqu'à représenter des caractéristiques uniques propres à l'objet. [16]

II.4.4.2 L'importance des réseaux de neurones convolutionnels (CNN)

- **Imagerie médicale :** Les CNN peuvent examiner des milliers de rapports pathologiques pour détecter la présence ou l'absence de cellules cancéreuses sur des images.

- **Traitement audio :** La détection de mots-clés peut être utilisée sur n'importe quel dispositif doté d'un microphone, de manière à détecter la prononciation d'un mot ou d'une phrase spécifique (« Dis Siri »). Les CNN peuvent apprendre et détecter des mots clés avec précision et ignorer le reste, quel que soit l'environnement.
- **Détection d'objets :** La conduite autonome s'appuie sur les CNN pour détecter les panneaux ou autres objets, et prendre des décisions en fonction de la sortie.
- **Génération de données synthétiques :** Grâce aux réseaux antagonistes génératifs (GAN) de nouvelles images peuvent être produites et utilisées dans des applications de Deep Learning comme la reconnaissance faciale ou la conduite autonome.[16]

II.4.4.3 Architecture d'un CNN

Les réseaux de neurones convolutionnels généralement composés des couches suivantes :

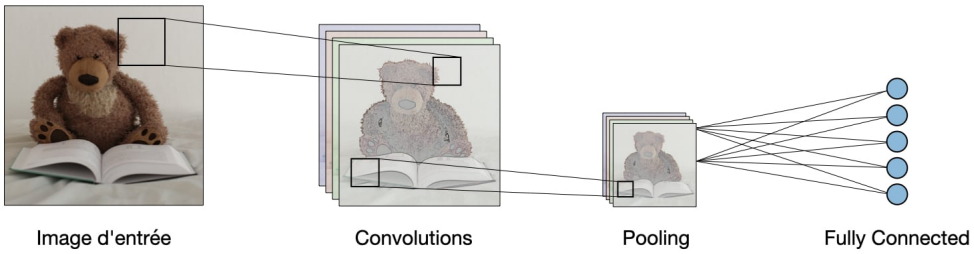


FIGURE II.16 – Couche d'un CNN

La couche convolutionnelle et la couche de pooling peuvent être ajustées en utilisant des paramètres qui sont décrites dans les sections suivantes.[17]

II.4.4.4 Types de couche

- **Couche convolutionnelle (CONV) :** La couche convolutionnelle (en anglais convolution layer) (CONV) utilise des filtres qui scannent l'entrée **I** suivant ses dimensions en effectuant des opérations de convolution. Elle peut être réglée en ajustant la taille du filtre **F** et le stride **S**. La sortie **O** de cette opération est appelée feature map ou aussi activation map.

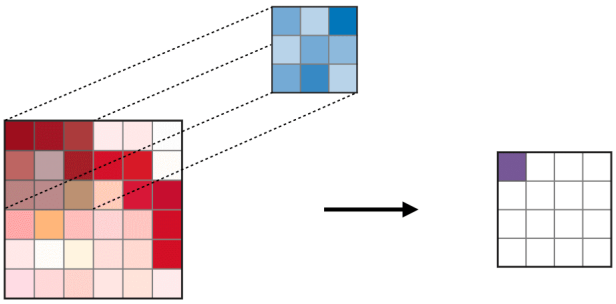


FIGURE II.17 – Couche convolutionnelle

✦ **Pooling (POOL) :** La couche de pooling (en anglais pooling layer) (POOL) est une opération de sous-échantillonnage typiquement appliquée après une couche convolutionnelle. En particulier, les types de pooling les plus populaires sont le max et l'average pooling, où les valeurs maximales et moyennes sont prises, respectivement.

✻ **Max pooling :** Son but est a Chaque opération de pooling sélectionne la valeur maximale de la surface.

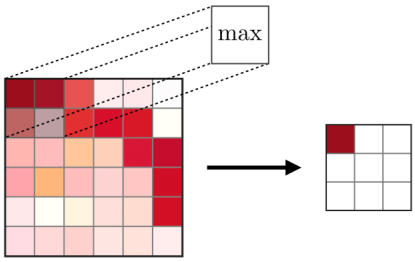


FIGURE II.18 – Max pooling

✻ **Average pooling :** Son but est a Chaque opération de pooling sélectionne la valeur moyenne de la surface.

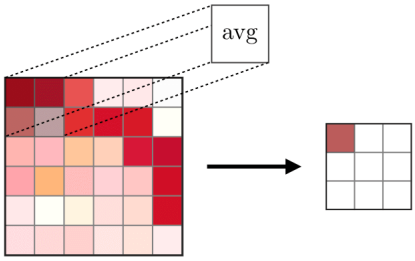


FIGURE II.19 – Average pooling

✦ **Fully Connected (FC) :** La couche de fully connected s'applique sur une entrée préalablement aplatie. Peuvent être utilisées pour optimiser des objectifs tels que les scores de classe.[17]

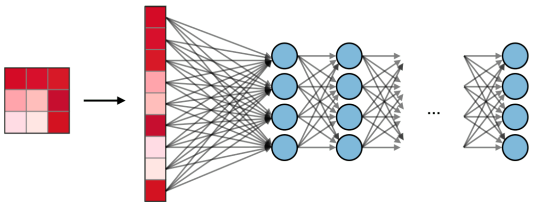


FIGURE II.20 – fully connected

II.5 La difference entre le ML et DP

Aspect	**Machine Learning (ML)**	**Deep Learning (DL)**
Architecture	Souvent basée sur des modèles d'apprentissage supervisé ou non supervisé, comme les arbres de décision, les SVM, les réseaux de neurones peu profonds, etc.	Basée sur des réseaux de neurones profonds composés de nombreuses couches cachées.
Représentation des données	Les caractéristiques des données sont souvent extraites manuellement ou avec des techniques de sélection de caractéristiques.	Les réseaux de neurones apprennent automatiquement les caractéristiques des données à partir des données brutes, en évitant souvent la nécessité d'une extraction de caractéristiques manuelle.
Taille des données	Convient généralement à des ensembles de données plus petits et moins complexes.	Convient à des ensembles de données massifs, notamment des données non structurées telles que des images, des textes et des vidéos.
Complexité des modèles	Moins de complexité dans la structure des modèles, moins de couches cachées et moins de paramètres à apprendre.	Les modèles peuvent être extrêmement complexes, avec de nombreuses couches cachées et des millions de paramètres, ce qui nécessite souvent un grand nombre de données pour l'entraînement et des ressources informatiques puissantes pour l'inférence.
Exigences en matière de calcul	Moins exigeant en termes de puissance de calcul et de ressources matérielles.	Plus exigeant en termes de puissance de calcul, nécessitant souvent l'utilisation de GPU ou de TPU pour l'entraînement et l'inférence sur des ensembles de données massifs.
Performance	Peut être moins performant sur des tâches complexes nécessitant une représentation de données hautement non linéaire.	Peut être plus performant sur des tâches complexes impliquant des données non structurées telles que la reconnaissance d'images, le traitement du langage naturel et la reconnaissance vocale.
Interprétabilité	Souvent plus interprétable car les modèles sont moins complexes et les caractéristiques sont souvent explicites.	Souvent moins interprétable en raison de la complexité des modèles et de la nature implicite des caractéristiques apprises.

TABLE II.3 – Comparaison entre le Machine Learning (ML) et le Deep Learning (DL)

II.6 Conclusion

Dans ce chapitre, nous avons exploré l'utilisation des techniques d'apprentissage automatique et d'apprentissage profond pour l'analyse et la classification de données médicales, en particulier dans le cadre de la prédiction et du diagnostic du cancer du sein. Ces approches ont montré leur grande efficacité pour traiter des ensembles de données complexes et volumineuses issus d'examens cliniques, d'imagerie médicale et d'autres sources pertinentes.

Les algorithmes d'apprentissage automatique classiques tels que les arbres de décision, les forêts aléatoires ou les machines à vecteurs de support se sont révélés très utiles pour réaliser des classifications précises à partir de caractéristiques extraites manuellement des données. Leur capacité à détecter des motifs complexes et à généraliser à partir d'exemples les rend particulièrement adaptés à cette tâche de diagnostic assisté.

D'un autre côté, les réseaux de neurones profonds et autres techniques d'apprentissage profond apportent une puissance supplémentaire en étant capables d'extraire et d'apprendre automatiquement les caractéristiques discriminantes directement à partir des données brutes,comme les images histopathologiques du sein. Malgré leur plus grande complexité, ces modèles ont permis d'atteindre des performances très élevées, proches ou dépassant l'expertise humaine pour certaines tâches spécifiques.

Cependant, l'adoption de ces méthodes intelligentes dans un cadre clinique soulève des défis importants en termes d'explicabilité, de robustesse, de sécurité et d'éthique qui devront être soigneusement pris en compte. Une utilisation responsable, en complément de l'expertise médicale humaine existante, sera la clé pour tirer la meilleur parti de ces technologies prometteuses au service de la santé des patients.

ChapitreIII

Etat de L'Art

III.1 Introduction

Au cours de ce chapitre, nous examinons les recherches pertinentes sur les techniques, méthodes et applications les plus couramment utilisées pour la prédiction, le diagnostic et le traitement des maladies néoplasiques du sein, les datasets les plus cités et l'utilisation des réseaux de neurones convolutifs(CNN) ainsi les algorithmes de machine learning tels que la régression logistique, les arbres de décision et les SVMs.

Nous traiterons cette problématique en examinant d'une manière approfondie plusieurs articles pertinents qui ont contribué significativement à ce domaine. Ces articles présenteront leurs méthodologies, leurs approches et les résultats obtenus, nous permettant d'explorer les avancées récentes et les défis associés à l'utilisation des CNN.Nous pourrons comprendre les différentes architectures et techniques utilisées, ensembles de données utilisés pour l'entraînement et la validation, ainsi que les performances obtenues par ces modèles.

L'objectif de ce chapitre d'état de l'art est de fournir une vision approfondie et complète de l'utilisation d'algorithmes du ML, CNN ans la prédiction et la classification. En examinant les différentes approches et les résultats obtenus, nous espérons contribuer à l'enrichissement des connaissances et à l'avancement de l'IA.

III.2 Approche classique

Pour savoir si une personne est atteint d'un cancer du sein malin ou bénin, les experts de la santé utilise des méthodes traditionnelles qui reposent sur des examens physiques complets, en plus d'un examen des seins, afin de mieux comprendre la cause des symptômes. Les tests les plus réputés pour aider à diagnostiquer sont :

III.2.1 Examen clinique des seins

Un examen clinique des seins est un examen médical comprenant l'inspection visuelle et la palpation des seins, y compris la région des aisselles.L'inspection visuelle des seins permet au médecin de rechercher une [31] :

- ✦ Anomalie de la forme du sein.
- ✦ Modification de la coloration de la peau.
- ✦ Anomalie au niveau du mamelon.

- Les défis :

- ✜ La sensibilité est faible.
- ✜ Elle ne détermine pas la malignité.
- ✜ Dangereux en ce qui concerne les biopsies supplémentaires requises.

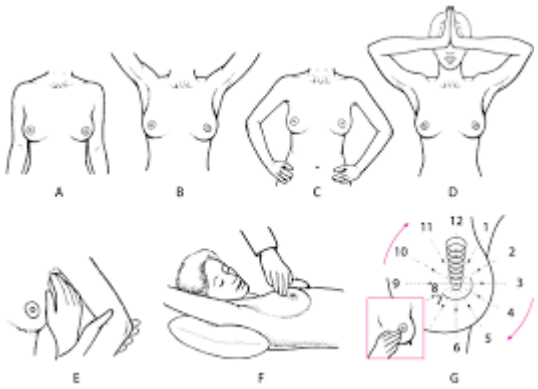


FIGURE III.1 – Auto-palpation

III.2.2 Mammographie

La mammographie est une modalité d'imagerie médicale mettant en œuvre la radiographie des seins, en vue de dépister des anomalies et le plus souvent, de dépister des cancers du sein. Elle permet d'obtenir des images des tissus intérieur et ainsi de détecter au plus tôt d'éventuelles anomalies, notamment des nodules, qui peuvent être signes d'un cancer du sein. L'image est enregistrée dans un fichier électronique plutôt que dans un film lors d'une mammographie numérique.[32]

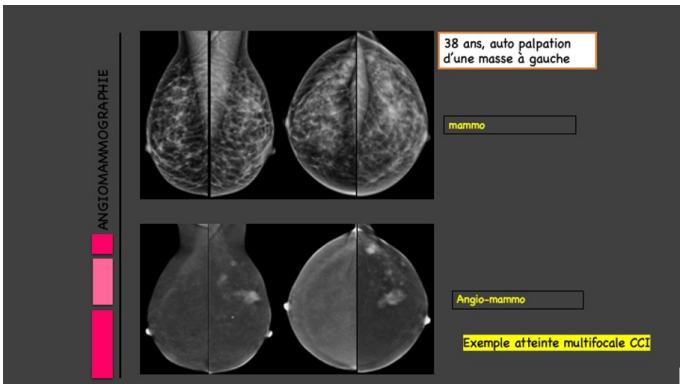


FIGURE III.2 – Mammographie

- Les défis :

- Risque pour les jeunes femmes de moins de 40 ans.
- Plage dynamique limitée, faible contraste et image granuleuse.
- La densité mammaire affecte la sensibilité et la spécificité.

III.2.3 IRM mammaire

L'imagerie par résonance magnétique du sein (IRM) est un examen radiologique complémentaire destiné à affiner le diagnostic. On l'utilise aussi pour la surveillance des femmes à haut risque familial de cancer du sein. L'appareil d'IRM ressemble à un gros cylindre que traverse un lit mobile. Toutes les anomalies détectées par une IRM mammaire ne sont pas cancéreuses. Très fréquemment, une échographie est programmée après l'IRM dans le but de controler l'image vue en IRM. Selon l'anomalie détectée et son apparence suspecte, le médecin décide ou non de réaliser d'autres examens pour le diagnostique du patient. [33]

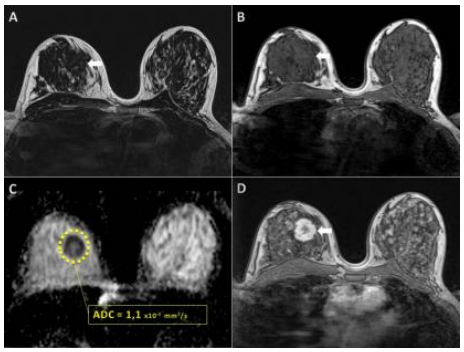


FIGURE III.3 – IRM

- Les défis :

- La précision est limitée.
- Temps d'imagerie lent.
- Seulement recommandé pour le dépistage des femmes à risque élevé.

III.2.4 Échographie mammaire

L'échographie mammaire est un examen d'imagerie qui utilise les ultrasons pour obtenir des clichés de l'intérieur de la glande mammaire. Elle est souvent pratiquée en association d'une mammographie dans le cadre du dépistage et du suivi du cancer du sein.[34]

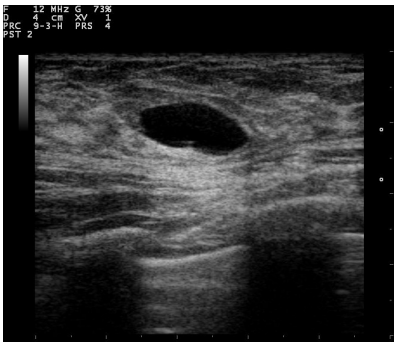


FIGURE III.4 – echographie

- Les défis :

- ✜ Difficulté a observer les tissus profonds ou denses .
- ✜ La patiente sera mal a l'aise quand une pression est exercée sur les seins.
- ✜ Difficulté a observer certaines lésions surtout chez les femmes ayant des seins denses.

III.2.5 PET Scan

Le Pet Scan (ou TEP, Tomographie par émission de positons) détecte les cellules cancéreuses grâce à un examen rapide. En injectant un produit radioactif à base de glucose, une caméra spéciale couplée avec un scanner suit son déplacement dans l'organisme. Les cellules cancéreuses, étant très actives, consomment beaucoup plus de glucose que les cellules normales. Cette concentration élevée de glucose dans les tissus tumoraux est mise en évidence par le TEP Scan. [35]

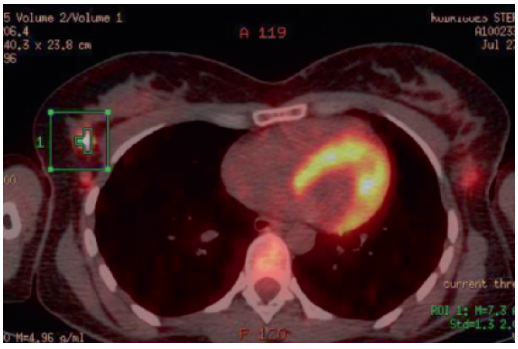


FIGURE III.5 – PET scan

- Les défis :

- ✜ Couts élevés .
- ✜ Temps d'imagerie lent.

III.3 Approches intelligentes

Récemment, les méthodes avancées de l'apprentissage automatique, notamment les réseaux de neurones à convolution en couches profondes, ont connu un succès remarquable dans l'analyse des images médicales. Cette réussite découle principalement de la capacité des multiples couches de ces réseaux à apprendre des représentations significatives à partir des données, ce qui conduit à des améliorations dans la capacité à identifier et caractériser les entités dans les images médicales. Bien que ces avancées soient prometteuses, trés couteuses et nécessite l'ajustement les hyperparamètres et la structure des modèles.

III.3.1 Modèles et méthodes existants

Pour la classification, la prédiction, la segmentation et le diagnostic du cancer du sein il existe plusieurs approche et modèles CNN qui ont été proposées et appliquées avec succès sur différents ensembles de données d'imagerie médicale.

III.3.1.1 Ségmentation

La segmentation d'image est une tâche de vision par ordinateur qui consiste à diviser une image particulière en plusieurs segments, chaque segment représentant un objet ou une région de l'image. Cette tâche est importante pour des applications telles que la détection d'objets, la reconnaissance d'images et la conduite autonome. [36]
Parmi les modèles populaires pour cette tâche, on peut citer U-Net, Mask R-CNN, DeepLabV3+, Attention U-Net et FractalNet.
Un chercheur a permis de réduire la complexité d'un modèle tout en gardant une bonne précision de segmentation sur plusieurs ensembles de données en utilisant l'architecture U-Net[36]. Cette variante compacte appelée TernausNet parvient à diminuer considérablement le nombre de paramètres par rapport à U-Net classique.[37]

III.3.1.2 Détection

Il existe deux principales méthodes utilisées pour la détection des noyaux. La première méthode est basée sur la détection, qui nécessite une étape de segmentation préalable, comme discuté dans[38].
Les modèles populaires pour cette approche incluent Faster R-CNN, YOLO, SSD. En revanche, la deuxième méthode repose sur l'estimation de la densité pour détecter les noyaux sans utiliser de segmentation, comme expliqué dans la référence[39]. Dans cette approche, un cadre basé sur un CNN avec un pool maximal super-visé est formé pour détecter la région de pixels de cellule, préalablement pré-sélectionnée à l'aide d'un SVM. Le modèle de régression basé sur le CNN est ensuite utilisé pour détecter et compter les noyaux. Ce modèle de réseau de régression neurale entièrement convolutif est capable de cartographier la densité pour une image d'entrée de taille arbitraire.

III.3.1.3 Classification

Des chercheurs ont exploré plusieurs approches pour la classification des cellules dans le cadre du diagnostic du cancer du sein. Ils ont analysé des caractéristiques telles que la forme, la texture et la taille des noyaux dans les images. Certains ont utilisé des CNN comme VGGNet, ResNet, Inception pour distinguer les cellules mitotiques des cellules

non mitotiques en se basant sur leurs propriétés visuelles. D'autres ont développé des méthodes de classification en utilisant des informations sur l'intensité, la morphologie et la texture des noyaux avec des classifieurs comme les machines à vecteurs de support (SVM), les forêts aléatoires ou les k-plus proches voisins. Plus récemment, les approches basées sur l'apprentissage profond, comme SVM, AdaBoost, les réseaux de neurones convolutifs profonds (DCNN) et les architectures attentionnelles ont montré des résultats prometteurs pour la classification des images pathologiques.

III.4 Datasets

Il existe plusieurs jeux de données qui sont privé ou public, la plupart d'entre eux sont disponibles sous deux formats différents :

- Fichiers CSV (Comma Separated Values).
- Images (jpg, pgm, png, DICOM et jpeg).

Le dataset le plus populaire et le plus utilisé est celui de de Breast Cancer Wisconsin (Diagnostic) Il contient 569 échantillons d'images dont 212 maligne et 357 bénine. Le résumé des jeux de données sur le cancer du sein les plus citées est présenté dans le tableau suivant :

Dataset	**Size**	**Classes**	**Format**	**Modalité**
MIAS	322	N, AB	Pgm	DM
DDSM	55,890	N, AB	Jpeg	DM
Breast Cancer Wisconsin (Diagnostic)	569	B, M	CSV	DM
Breast Histopathology	277,524	IDC (-), IDC (+)	Png	DM
NKI Breast Cancer Data	295	1, 0	CSV	Gene
Breast Cancer Proteomes	77	PAM50, mRNA	CSV	Gene
Incidence of breast cancer	150	B, M, In situ	CSV	Gene
IRMA	14k	-	jpg	-
UM : University of Michigan	-	-	-	-
UC : University of Chicago	-	-	-	-

Table III.1 – Breast cancer datasets

III.5 Travaux connexes

Pour la détection précoce, le diagnostic et le traitement des maladies, les images médicales sont principalement interprétées par des experts humains tels que des radiologues et des médecins. Cependant, en raison des grandes différences de pathologie et de la fatigue potentielle des experts humains, les chercheurs et les cliniciens ont récemment commencé à bénéficier des interventions assistées par ordinateur. voici quelques travaux basés sur le ML et DL :

- **Article 1 : Breast cancer prediction using different machine learning methods applying multi factors**[23]

  Cette étude compare diverses techniques d'apprentissage automatique pour créer un modèle de prédiction du risque de cancer du sein. Basé sur un échantillon de 810 individus, incluant 115 patients atteints de cancer et 695 personnes en bonne santé, 45 attributs ont été sélectionnés par des experts parmi 85 disponibles, couvrant des facteurs génétiques, biochimiques, biomarqueurs, démographiques et pathologiques. Le modèle de forêt aléatoire (RF) se distingue avec une précision de 99,26 % et une AUC de 99 %. Les variables les plus influentes sur le risque de cancer du sein sont la pathologie, les biomarqueurs, la biochimie, les gènes et les facteurs démographiques. Cette approche montre qu'une prise en compte exhaustive des caractéristiques permet de développer un modèle de prédiction précis, tel que celui obtenu ici avec une précision de 99,3 %.

- **Article 2 : Machine Learning Classification Techniques for Breast Cancer Diagnosis**[24]

  L'article examine l'utilisation de systèmes de détection assistée par ordinateur (CAD) dans le diagnostic précis du cancer du sein, une maladie courante et l'une des principales causes de décès par cancer chez les femmes. Les CAD, exploitant les techniques d'apprentissage automatique, peuvent contribuer à une détection précoce, améliorant ainsi les chances de survie. L'étude compare l'efficacité de différentes méthodes d'apprentissage automatique, notamment les machines à vecteurs de support, les réseaux neuronaux artificiels et le Naïve Bayes, en utilisant l'ensemble de données Wisconsin Diagnostic Breast Cancer (WDBC). Une approche hybride est proposée, combinant la réduction de la dimensionnalité des caractéristiques avec l'analyse discriminante linéaire (LDA), suivie de l'application des données réduites aux machines à vecteurs de support. Cette méthode a démontré une précision de 98,82 %, une sensibilité de 98,41 %, une spécificité de 99,07 % et une excellente aire sous la courbe ROC de 0,9994.

- **Article 3 : reast Cancer Detection and Diagnosis Using Mammographic Data : Systematic Review**[25]

  Cette revue systématique évalue les différentes méthodes d'intelligence artificielle utilisées pour la détection et le diagnostic du cancer du sein à partir de données mammographiques. Les auteurs ont analysé 92 études publiées entre 2010 et 2019. La majorité des études ont utilisé des réseaux de neurones convolutionnels (CNN) comme technique d'apprentissage profond pour analyser les images mammographiques. Les CNN ont montré de meilleures performances que les méthodes tradi-

tionnelles pour la classification des masses mammaires et la détection des microcalcifications. Plusieurs études ont combiné les CNN avec d'autres algorithmes comme les machines à vecteurs de support (SVM) ou les forêts d'arbres décisionnels, améliorant encore les résultats. L'utilisation de techniques de transfert d'apprentissage à partir de réseaux pré-entraînés sur de grands ensembles de données a également permis d'augmenter les performances.Les datasets publics les plus utilisés étaient INbreast, DDSM, CBIS-DDSM et MIAS. Cependant, les petites tailles d'échantillons et les biais de sélection des datasets restent des défis majeurs.

En conclusion, l'apprentissage profond, en particulier les CNN, semble très prometteur pour améliorer la détection et le diagnostic du cancer du sein à partir d'images mammographiques, mais des recherches supplémentaires sur de grands datasets sont nécessaires pour une application clinique fiable.

- **Article 4 : Performance of a Breast Cancer Detection AI Algorithm Using the Personal Performance in Mammographic Screening Scheme**[26]

  Cette étude évalue les performances d'un algorithme d'intelligence artificielle (IA) pour la détection du cancer du sein sur des mammographies de dépistage.

  L'algorithme d'IA a été entraîné sur un large ensemble de données mammographiques et de scores de rappel du programme de dépistage Personnel Performance in Mammographic Screening (PERFORMS).

  Les performances de l'algorithme ont été testées sur un ensemble indépendant de 88 633 mammographies de dépistage en double lecture.

  L'algorithme d'IA a obtenu une sensibilité de 84,7% et une spécificité de 92,4% pour la détection des cancers, dépassant légèrement les performances des radiologistes (83,7% de sensibilité et 91,6% de spécificité).

  Combiné avec la double lecture des radiologistes, l'algorithme a permis d'augmenter la sensibilité à 88,6%, tout en maintenant une spécificité de 89,8%.

  Cette étude démontre le potentiel de l'IA pour améliorer les performances de détection du cancer du sein en mammographie, en complément de l'interprétation par les radiologistes.

  Cependant, des études supplémentaires sur de grands ensembles de données sont nécessaires pour confirmer ces résultats prometteurs avant un déploiement clinique à grande échelle.

- **Article 5 : Breast Cancer Detection Using Deep Learning Technique Based On Ultrasound Image** [27]

  Cet article traite de l'utilisation de l'apprentissage profond pour améliorer la précision de la classification des types de cancer du sein à partir d'images échographiques. Le cancer du sein représente un défi diagnostique majeur en raison de la grande variabilité en termes de taille, forme, apparence et position des tumeurs, en particulier pour les cancers malins. Les auteurs proposent un système d'apprentissage profond en plusieurs étapes. Premièrement, un prétraitement des images est effectué pour améliorer leur qualité. Ensuite, la segmentation des régions d'intérêt est réalisée à l'aide de l'architecture U-Net. De nombreuses caractéristiques sont alors extraites des images segmentées en utilisant le réseau MobileNet. Enfin, ces caractéristiques alimentent un classifieur entraîné à différencier les types de cancer du sein. Les résultats obtenus sur un jeu de données d'images échographiques mammaires sont très

prometteurs, avec une précision de classification atteignant 99,29%, dépassant les performances des travaux précédents sur le même sujet.

Cette approche par apprentissage profond semble particulièrement efficace pour faire face au manque d'uniformité des cancers du sein en termes de forme, taille et localisation. En combinant prétraitement d'image, segmentation précise et extraction de caractéristiques pertinentes, le système proposé parvient à une très grande justesse de diagnostic sur ces images médicales complexes.

- **Article 6 : Three-Class Mammogram Classification Based on Descriptive CNN Features** [28]

Dans cette article les chercheurs ont introduit des techniques de classification multi-classes basées sur CNN-CT (ComputerTomography)pour classer les mammographies des ensembles de données IRMA en normales, bénignes et malignes. La fusion des caractéristiques CNN et des caractéristiques les plus descriptives avec des ondelettes a bien fonctionné et a atteint une précision de 83,74% pour le classificateur SVM.

- **Article 7 : Performance comparison of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignante** [29]

Cet article parle de deux approches différentes pour la classification d'images de cancer du sein en bénignes ou malignes à partir d'images IRM :

La première approche est basée sur la segmentation d'images et l'extraction manuelle de 38 caractéristiques de différentes catégories (texture, taille, morphologie, etc.) après segmentation.

La seconde approche utilise un réseau de neurones convolutionnels (CNN) pré-entraîné AlexNet. Des vecteurs de caractéristiques de 4096 dimensions sont extraits des couches complètement connectées, puis réduits à 518 caractéristiques pertinentes. L'étude a été réalisée sur un ensemble de 640 images IRM, dont 191 bénignes et 449 malignes. Les performances ont été évaluées avec un classifieur LDA et une validation croisée, en testant séparément les caractéristiques issues de la segmentation (38), les caractéristiques CNN (518) et la fusion des deux (556 caractéristiques). Les résultats ont montré que l'approche par caractéristiques CNN permettait d'atteindre une précision de 0,88, contre 0,76 pour la segmentation. Cependant, la fusion des deux ensembles de caractéristiques a permis d'obtenir la meilleure précision de 0,91 pour la classification bénin/maligne sur ces images IRM. Cette étude démontre ainsi l'intérêt de combiner les techniques d'extraction de caractéristiques manuelles et automatiques par apprentissage profond pour améliorer les performances de diagnostic assisté par ordinateur en imagerie médicale.

- **Article 8 : Classification of breast cancer histology images using Convolutional Neural Networks** [30]

Le cancer du sein est l'une des principales causes de décès par cancer dans le monde. Le diagnostic des tissus de biopsie avec des images colorées à l'hématoxyline et à l'éosine est complexe, et les spécialistes sont souvent en désaccord sur le diagnostic final. Les systèmes de diagnostic assisté par ordinateur contribuent à réduire les coûts et à augmenter l'efficacité de ce processus. Les approches de classification conventionnelles reposent sur des méthodes d'extraction de caractéristiques conçues

pour un problème spécifique basé sur la connaissance du domaine. Pour surmonter les nombreuses difficultés des approches basées sur les caractéristiques, les méthodes d'apprentissage profond deviennent des alternatives importantes. Une méthode pour la classification des images de biopsie du sein colorées à l'hématoxyline et à l'éosine utilisant des réseaux de neurones convolutionnels (CNN) est proposée. Les images sont classées en quatre catégories : tissu normal, lésion bénigne, carcinome in situ et carcinome invasif, et en deux catégories : carcinome et non-carcinome. L'architecture du réseau est conçue pour récupérer des informations à différentes échelles, incluant à la fois les noyaux et l'organisation globale des tissus. Cette conception permet l'extension du système proposé aux images histologiques de lames entières. Les caractéristiques extraites par le CNN sont également utilisées pour entraîner un classificateur à vecteurs de support (SVM). Des précisions de 77,8% pour les quatre classes et de 83,3% pour carcinome/non-carcinome sont atteintes. La sensibilité de notre méthode pour les cas de cancer est de 95,6%.

III.6 Etude et comparaison

Nous réalisons une étude comparatives utilisant les approches suivantes selon les 7 facteurs dans les tableaux ci-dessous :

- ✧ **Titre :** Représente le titre de l'article et les noms des auteurs.
- ✧ **Dataset :** Désigne les Dataset utilisées pour l'implémentation de la proche proposée.
- ✧ **Approche :** Désigne l'approche de chaque article.
- ✧ **Résultats :** Les résultats obtenue par chaque approche.
- ✧ **Avantages :** Les avantages de l'approche abordée.
- ✧ **Inconvénients :** Les inconvénients de l'approche abordée.

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
Breast cancer prediction using different machine learning methods applying multi factors (Elham Nazari, Hamid Naderi, Mahla Tabadkani, Reza Aref Nezhad, Amir Hossein Farzin, Mohammad Dashtiahangar, Majid Khazaei, Gordon A. Ferns, Amin Mehrabian, Hamed Tabesh and Amir Avan)	un jeu de données qui contient 810 instances (115 patients atteints du cancer et 695 individus en bonne santé) et 45 attributs	une comparaison de 13 modèles de ML pour développer un modèle de prédiction de risque de cancer du sein.	Les meilleures performances ont été obtenues par le Random Forest (RF) avec une précision de 99%, une précision de 99.26% et une AUC de 99%.	Dans la prédiction du risque de cancer du sein, l'approche RF a démontré une grande précision, atteignant une précision de 99,3%.	Les modèles dépendent de la qualité des données utilisées pour l'entraînement

Table III.2 – Étude du premier article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
Machine Learning Classification Techniques for Breast Cancer Diagnosis (David A.Omondiagbe ,Shanmugam Veeramani, Amandeep S.Sidhu)	le jeu de données winsconsin diagnostic breast cancer (WDBC)	les techeniques utilisées dand cete étude SVM, ANN, Naive Bayes classifier,Correlation based Feature Selection,Recursive Feature Elimination,Principal Component Analysis,Linear Discriminant Analysis	la meilleure approche qui a obtenu des meilleurs performance est SVM avec une précision de 98.82% une sensibilité de 98,41 %, une spécificité de 99,07 %	Démontre que la sélection et l'extraction de caractéristique pertinentes put améliorer le diagnostic des tumeurs bénignes malignes , Propose une approche hybride combinant les avantages de la réduction de dimension et de l'apprentissage machine.	L'étude se concentre uniquement sur un ensemble de données (WDBC)

TABLE III.3 – Étude du Deuxième article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
Breast Cancer Detection and Diagnosis Using Mammographic Data : Systematic Review (PhD Syed Jamal Safdar Gardezi, PhD Ahmed Elazab, PhD Baiying Lei, PhD Tianfu Wang)	plusieurs base de données privées et publiques de la mammographie (DDSM, INbreast, CBIS-DDSM, MIAS)	Revue de différentes technique d'apprentissage automatqiue et profond pour la détection et la classification des masses et calcifications (SVM, ANN, KNN, CNN, RNN/LSTM, AlexNet, VGGNet, GoogLeNet, ResNet, etc.)	la meilleure approche de la classification est celle de Shams et al et de Levy and Jain avec une précision de 92,4 % pour levy, et d'AUC de shams (aire sous la courbe ROC) de 0,925 sur les datasets publics INbreast et DDSM en utilisant des vues multiples.	L'utilisation de techniques de transfert d'apprentissage améliore significativement les performances des modèles.	Le système dépend fortement de la disponibilité de données d'entraînement de haute qualité et en quantité suffisante.

TABLE III.4 – Étude du troisiéme article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
Performance of a Breast Cancer Detection AI Algorithm Using the Personal Performance in Mammographic Screening Scheme (Yan Chen PhD, Adnan G. Taib BMBS,Iain T.Darker PhD, Jonathan J.James FRCR)	Aucun dataset n'a été mentionné dans l'article, ils ont seulement indiqué que deux ensembles de tests PERFORMS qui ont été utilisés pour évaluer les performances de l'algorithme déjà entraîné	ils ont utilisé un algorithme d'IA commercialisé (Lunit INSIGHT MMG), En comparant les performances de 552 radiologue expert sur les mêmes ensembles de tests	L'IA a montré des performances équivalentes aux 552 lecteurs humains pour détecter le cancer du sein sur les ensembles de tests PERFORMS lorsque ses seuils étaient ajustés aux niveaux des lecteurs humains.	Permit d'évaluer les résultats de l'IA par rapport à un grand nombre de médecins et radiologues sur les mêmes situations cliniques. Adopte un modèle d'assurance qualité déjà en place.	Taille d'échantillon limitée. Spécificité de l'IA supérieure aux lecteurs humains avec le seuil de rappel recommandé par le développeur.

Table III.5 – Étude du Quatrième article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
Breast Cancer Detection Using Deep Learning Technique Based On Ultrasound Image (Abdulqader Mohammed, Mohammed Abdel Razek, Mohamed Eldosuky, Ahmed Sobhi)	base de donnée de 780 images sans masque et 1583 avec masque d'ultrasons mammaires provenant de l'hôpital Baheya en Égypte.	Prétraitement, Segmentation avec U-Net 3) Extraction de caractéristiques avec MobileNet et Classification avec VGG16	En combinant le prétraitement, la segmentation U-Net, l'extraction de caractéristiques MobileNet et la classification VGG16, le système proposé offre une précision maximale de 99,29%.	Haute précision supérieure aux travaux précédents Combinaison de diverses versions DL.	La taille relativement petite du dataset utilisé dans cetteétude peut limiter la généralisation des résultats

Table III.6 – Étude du Cinquième article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
Three-Class Mammogram Classification Based on Descriptive CNN Features (M. Mohsin Jadoon, Qianni Zhang, Ihsan Ul Haq, Sharjeel Butt, nd Adeel Jadoon)	base de données IRMA composée de 2796 patchs d'images mammographiques provenant de différentes sources (DDSM, MIAS, LLNL, RWTH)	Deux approches proposées : Réseau de neurones convolutionnel transformée en ondelettes discrète (CNN-DW) Réseau de neurones convolutionnel transformée en curvelettes (CNN-CT)	CNN-DW avec une précision de 81,83% (SVM) et 81,23% (SVM 10-fold cross-validation), CNN-CT : Précision de 83,74% (SVM) et 83,11% (SVM 10-fold cross-validation).	Utilise l'apprentissage profond afin de classer les individus en trois catégories, Utilise SVM au lieu de Softmax pour améliorer les performances .	Nécessite un grand ensemble de données, Les performances sont légèrement inférieures avec la validation croisée 10-fold SVM

TABLE III.7 – Étude du Sixième article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
(Performance comparison of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignante)	L'étude a été réalisée sur un ensemble de 640 images privés IRM,extraites par l'université de Chicago	Approche par segmentation,38 attributs répartis en 6 catégories ont été extraites aprés la segmentation des images en vue de la classification,pour l'approche CNN des régions de d'intéret ont été extraites et fournies en entrées au modéle AlexNet pré-entrainé .	Les performances ont été évaluées au moyen d'un classificateur LDA avec validation croisées, la précision respectives étaient de 0.88 pour la segmentation , 0.76 pour la CNN et 0.91 pour les caractéristiques fusionnées.	L'avantage de fusionner les caractéristiques de segmentation et du CNN est une amélioration significative de la précision de classification, grâce à une combinaison complémentaire des informations extraites des deux approches.	la complexité accrue de l'approche de fusion, nécessitant une gestion plus complexe des caractéristiques combinées et potentiellement une augmentation des temps de calcul

TABLE III.8 – Étude du Septième article

Titre	Dataset	Approche	Résultats	Avantages	Inconvénients
(Classification of breast cancer histology images using Convolutional Neural Networks)	l'étude a été realisée sur le jeu de données Breast Histology qui contient des images histologiques du sein de haute résolution, réparties en quatre catégories : tissu normal, lésion bénigne, carcinome in situ et carcinome invasif.	L'approche utilise des réseaux de neurones convolutionnels (CNN) pour classifier automatiquement les images de biopsies mammaires en quatre catégories et entraîne un classificateur SVM avec les caractéristiques extraites par le CNN.	Différentes études ont utilisé des réseaux de neurones convolutionnels pour classifier les images histologiques du sein : Cruz-Roa et al ont obtenu une sensibilité de 79,6% pour la détection du carcinome invasif par fragments, tandis que cette étude a atteint environ 84% de précision pour la classification des tumeurs bénignes ou malignes à un grossissement de 200×.	Utilisation efficace des réseaux de neurones convolutionnels (CNN) pour classifier automatiquement les images histologiques du sein en plusieurs catégories.	La performance du CNN peut être influencée par la qualité et la quantité des données d'entraînement disponibles.

TABLE III.9 – Étude du huitième article

III.7 La conclusion

Dans ce chapitre, nous avons présenté un aperçu des principales approches, techniques et applications destinées à faciliter le diagnostic du cancer du sein.Nous avons mis en lumière les défis inhérents à chacune de ces techniques et les avons comparées.Les approches intelligentes basées sur des algorithmes d'apprentissage profond (Deep Learning) se sont révélées particulièrement puissantes pour le diagnostic du cancer du sein.Elles offrent une solution intelligente permettant de créer un pont entre l'informatique et la médecine. Ces algorithmes avancés établissent un lien concret entre ces deux domaines.

ChapitreIV

Conceptions et Résultats

IV.1 Introduction

Après avoir posé les bases théoriques nécessaires à la compréhension du cancer du sein, de ses techniques de dépistage, ainsi que des concepts liés à l'apprentissage automatique et aux réseaux de neurones artificiels, nous abordons dans ce chapitre les aspects pratiques liés à la conception et à la réalisation de notre système d'aide a la prédiction et diagnostic.

Nous présenterons dans un premier temps les choix méthodologiques effectués pour la prédiction et la classification du cancer de sein. Nous justifierons le recours aux techniques retenues, en mettant en évidence leurs forces et faiblesses au regard de la problématique traitée.

Ensuite, nous exposerons le processus de sélection et de préparation du jeu de données utilisé pour l'entraînement et l'évaluation de nos modèles. Les critères ayant conduit au choix de ce dataset de référence seront explicités.

Enfin, nous décrirons l'environnement et les outils de développement mis en œuvre pour l'implémentation de notre solution. Le choix de ces différents éléments technologiques sera motivé en fonction des contraintes et objectifs du projet.

IV.2 Les outils de développement

IV.2.1 Outils matériels

Pour la réalisation de notre systéme de prédiction et de classification du cancer du sein, nous avons utilisé une machine qui offre des performances acceptable pour le machine learning dont voici ses caractéristiques :

Composant	Spécifications
CPU	Intel(R) Core(TM) i7-4710HQ CPU @ 2.50GHz
RAM	8 Go DDR4
GPU	NVIDIA GeForce GTX 850M 2gb

Table IV.1 – Spécifications Matérielles

Concernant le deep learning (vision par ordinateur) nous avons opté pour Google Colab avec une version pro qui offre plus de RAM et de CPU ainsi un accès totale au GPU et au TPU 24h/24h et une exécution en arrière-plan, Les notebooks Colab Pro continuent de s'exécuter même si vous fermez votre navigateur ou que vous vous déconnectez, ce qui est idéal pour les tâches longues.

IV.2.2 Outils logiciels

IV.2.2.1 Environnement

- ✯ **Anaconda :** Est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique).[41]
- ✯ **Jupyter-notebook :** Le bloc-notes Jupyter est un outil polyvalent qui facilite la collaboration et la communication des résultats d'analyse de données. Il s'adresse aux scientifiques des données, aux chercheurs, aux étudiants et à toute personne souhaitant explorer et comprendre des données de manière interactive.
- ✯ **Google colab :** Est un environnement de développement interactif et puissant pour le langage Python, accessible depuis n'importe quel appareil avec une connexion Internet. Il est particulièrement adapté aux besoins des data scientists, des chercheurs et des étudiants en Data science.

IV.2.2.2 Framework

- ✯ **TensorFlow :** Une plateforme open source de deep learning créée par Google. Equipée d'une API Python, elle propose une multitude d'outils pour entrainer et optimiser des réseaux de neurones artificiels.[42]
- ✯ **Keras :** Keras est une API de réseau neurone développée en Python. Il s'agit d'une bibliothèque open source qui s'exécute sur des frameworks tels que Theano et TensorFlow.Il est rédigé et maintenu par Francis Chollet, un membre de l'équipe Google Brain.

- **Taipy :** Taipy est une bibliothèque Python open-source qui permet de créer rapidement des dashbords interactifs et des applications back-end. Elle offre des outils pour gérer les données et réaliser des analyses "what-if", ainsi qu'un éditeur graphique, Taipy Studio, pour simplifier la configuration des flux de données.

IV.2.2.3 Bibliothèques utilisées

- **Numpy :** NumPy est une bibliothèque Python très populaire qui est principalement utilisée pour effectuer des calculs mathématiques et scientifiques. Elle offre de nombreuses fonctionnalités et outils qui peuvent être utiles pour les projets de Data Science. [42]
- **Pandas :** Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse de données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.[43]
- **Matplotlib :**Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous forme de graphiques5. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy. Elle fournit également une API orientée objet, permettant d'intégrer des graphiques dans des applications, utilisant des outils d'interface graphique polyvalents tels que Tkinter, wxPython, Qt ou GTK.[44]
- **Scikit-learn :** Est une bibliothèque open source de machine learning en Python fournissant de nombreux algorithmes pour la modélisation et l'extraction de données. Elle offre une interface simple et performante pour les tâches courantes d'apprentissage automatique comme la classification, la régression et le clustering.
- **OpenCV (Open Source Computer Vision Library) :** Est une bibliothèque logicielle libre dédiée à la vision par ordinateur et au traitement d'images. Elle fournit une infrastructure pour les applications de vision artificielle avec de nombreux algorithmes de traitement d'images et de vidéos.

IV.2.2.4 Langage de programmation

- **Python :** Est un langage de programmation interprété, orienté objet et multiplateformes. Conçu pour favoriser la productivité des développeurs, Python se caractérise par sa syntaxe claire et concise, sa philosophie de programmation minimaliste et son type dynamique.

FIGURE IV.1 – Language de programmation python

Partie Prédiction

IV.3 Choix du dataset

Pour réaliser notre application de prédiction du cancer du sein, nous avons opté pour le dataset **Breast Cancer Winsconsin**[45] qui est adapté pour obtenir une bonne prédiction, grâce aux features pertinentes et riches qu'il fournit. ses features sont calculées à partir d'une image numérisée d'une aiguille fine aspirer (FNA) d'une masse mammaire. Ils décrivent les caractéristiques des noyaux de cellules présents sur l'image.

IV.3.1 Caractéristiques du dataset

Nombre d'instances : 569

Nombre d'attributs : 30 attributs numériques prédictifs et la classe.

Informations sur les attributs :

- rayon (moyenne des distances du centre aux points sur le périmètre).
- texture (écart-type des valeurs en niveaux de gris).
- périmètre.
- surface.
- douceur (variation locale des longueurs de rayon).
- compacité ($\frac{\text{périmètre}^2}{\text{surface}} - 1.0$).
- concavité (gravité des portions concaves du contour).
- points concaves (nombre de portions concaves du contour).
- symétrie.
- dimension fractale ("approximation de la côte" - 1).

Les moyennes, les erreurs-types et le "pire" ou le plus grand (moyenne des trois plus grandes valeurs) de ces caractéristiques ont été calculés pour chaque image, ce qui donne 30 caractéristiques. Par exemple, le champ 3 est le Rayon moyen, le champ 13 est le Rayon SE, le champ 23 est le Pire Rayon.
Classe :

- WDBC-Maligne
- WDBC-Bénigne

Synthèse statistique :

Caractéristique	Min	Max
Rayon (moyenne)	6.981	28.11
Texture (moyenne)	9.71	39.28
Périmètre (moyenne)	43.79	188.5
Surface (moyenne)	143.5	2501.0
Douceur (moyenne)	0.053	0.163
Compacité (moyenne)	0.019	0.345
Concavité (moyenne)	0.0	0.427
Points concaves (moyenne)	0.0	0.201
Symétrie (moyenne)	0.106	0.304
Dimension fractale (moyenne)	0.05	0.097
Rayon (erreur standard)	0.112	2.873
Texture (erreur standard)	0.36	4.885
Périmètre (erreur standard)	0.757	21.98
Surface (erreur standard)	6.802	542.2
Douceur (erreur standard)	0.002	0.031
Compacité (erreur standard)	0.002	0.135
Concavité (erreur standard)	0.0	0.396
Points concaves (erreur standard)	0.0	0.053
Symétrie (erreur standard)	0.008	0.079
Dimension fractale (erreur standard)	0.001	0.03
Rayon (pire)	7.93	36.04
Texture (pire)	12.02	49.54
Périmètre (pire)	50.41	251.2
Surface (pire)	185.2	4254.0
Douceur (pire)	0.071	0.223
Compacité (pire)	0.027	1.058
Concavité (pire)	0.0	1.252
Points concaves (pire)	0.0	0.291
Symétrie (pire)	0.156	0.664
Dimension fractale (pire)	0.055	0.208

TABLE IV.2 – Une synthèse statistique

Valeurs manquantes : Aucune

Distribution des classes : 212 - Maligne, 357 - Bénigne

Créateur : Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian

Donateur : Nick Street

Date : Novembre 1995

IV.4 Choix des méthodes

Dans le cadre de notre projet de prédiction du cancer du sein, nous avons choisie, exploré et comparé 6 algorithmes d'apprentissage supervisé :

IV.4.1 K-plus proches voisins (KNN)

Cet algorithme classe une nouvelle observation en fonction de la classe majoritaire de ses k plus proches voisins dans l'espace de caractéristiques. Son principal avantage est sa simplicité de mise en œuvre et d'interprétation.
Dans notre implémentation de l'algorithme KNN, le principal défi était de déterminer la valeur optimale du paramètre k, c'est-à-dire le nombre de voisins à considérer. Nous avons donc réalisé une recherche exhaustive de ce paramètre.

```
param_n = [3,4,5,6,9,11]
```

```
best_accuracy = 0
best_n = None
best_model = None


for n in param_n:

    knn = KNeighborsClassifier(n_neighbors=n)


    knn.fit(X_train, y_train)




    accuracy_knn = knn.score(X_test,y_test)

    if accuracy_knn > best_accuracy:
        best_accuracy = accuracy_knn
        best_n = n
        best_model = knn


print(f"Meilleur modèle avec n={best_n}, Exactitude: {best_accuracy:.4f}")
```

```
Meilleur modèle avec n=11, Exactitude: 0.9825
```

FIGURE IV.2 – implementation KNN

Pour chaque valeur de k testée, nous avons entraîné le modèle sur les données d'entrainement et évalué ses performances sur les données de test, afin de déterminer la valeur de k qui ajuste le mieux le modèle sur ces données. Le meilleur paramétre qui ajuste le mieux le modéle est n = 11.

IV.4.2 Arbre de décision

C'est un modèle de classification basé sur une série de règles de décision hiérarchisées qui subdivisent progressivement l'espace de caractéristiques. Ils sont faciles à interpréter mais peuvent souffrir d'un overfiting sur les données d'entraînement.
Pour appliquer cet algorithme, nous avons utilisé l'implémentation DecisionTreeClassifier de scikit-learn, afin d'obtenir les meilleures performances possibles, nous avons réalisé une recherche des meilleurs hyperparamètres par validation croisée.

```
param_max_depth = [None, 3, 5, 7, 9]
param_min_samples_split = [2, 5, 10]

best_accuracy_id3 = 0
best_params = {}

for max_depth in param_max_depth:
    for min_samples_split in param_min_samples_split:

        tree_clf = DecisionTreeClassifier(max_depth=max_depth, min_samples_split=min_samples_split, random_state=42)

        tree_clf.fit(X_train, y_train)

        y_pred_id3 = tree_clf.predict(X_test)

        accuracy_id3 = accuracy_score(y_test, y_pred_id3)
        if accuracy_id3 > best_accuracy_id3:
            best_accuracy_id3 = accuracy_id3
            best_params = {'max_depth': max_depth, 'min_samples_split': min_samples_split}

print("Meilleurs paramètres:", best_params)
print("Exactitude correspondante:", best_accuracy_id3)
```

```
Meilleurs paramètres: {'max_depth': 5, 'min_samples_split': 2}
Exactitude correspondante: 0.9649122807017544
```

FIGURE IV.3 – implementation Arbre de décision

Nous avons définie une liste de profondeur maximale de l'arbre ainsi le nombre minimum d'échantillons pour diviser le noead, pour chaque combinaison, nous avons entraîné un arbre de décision, évalué ses performances. la meilleurs hyperparamètres trouvé sont la profondeur maximale de l'arbre est de 5 et minimum d'échantillons requis pour qu'un nœud puisse être divisé est de 2.

IV.4.3 Machines à vecteurs de support (SVM)

L'approche de SVM cherche l'hyperplan séparateur optimal pour classer les observations en maximisant la marge entre les classes. Son extension non linéaire avec des noyaux le rend particulièrement puissant. Cependant, les SVM sont des boîtes noires difficiles à interpréter.

```
param_grid = {'kernel': ['linear', 'poly', 'rbf', 'sigmoid'], 'C': [0.1, 1, 10, 100]}
```

```
svm = SVC()
```

```
grid_search = GridSearchCV(estimator=svm, param_grid=param_grid, cv=5)
```

```
grid_search.fit(X_train,y_train)
```

```
▸ GridSearchCV
▸ estimator: SVC
    ▸ SVC
```

```
best_svm = grid_search.best_estimator_
best_svm
```

```
▾ SVC
SVC(C=1, kernel='linear')
```

```
accuracy_svm = best_svm.score(X_test,y_test)
```

```
accuracy_svm
```

```
0.9883040935672515
```

FIGURE IV.4 – implementation SVM

Nous avons exploré différentes kernels permettant de projeter les données dans un espace de plus grande dimension, ainsi que le paramètre de pénalité C contrôlant le compromis entre la maximisation de la marge et la minimisation des erreurs. La recherche exhaustive GridSearchCV a permis d'identifier la meilleure combinaison de noyau et de C, qui a ensuite été évaluée sur les données de test indépendantes. Le meilleur noyau trouvé est linear et C=1.

IV.4.4 Forêt aléatoire

Cet algorithme d'ensemble construit une multitude d'arbres de décision sur des sous-ensembles aléatoires des données et agrège leurs prédictions. Les forêts aléatoires sont réputées pour leur robustesse au sur-apprentissage et leur capacité à capturer des relations non linéaires complexes.

```
param_n_estimators = [10, 50, 100, 200]
param_max_depth = [None, 3, 5, 7]

best_accuracy_ran = 0
best_params_ran = {}

for n_estimators in param_n_estimators:
    for max_depth in param_max_depth:

        rf_clf = RandomForestClassifier(n_estimators=n_estimators, max_depth=max_depth, random_state=42)

        rf_clf.fit(X_train, y_train)

        y_pred = rf_clf.predict(X_test)

        accuracy_ran = accuracy_score(y_test, y_pred)

        if accuracy_ran > best_accuracy_ran:
            best_accuracy_ran = accuracy_ran
            best_params_ran = {'n_estimators': n_estimators, 'max_depth': max_depth}

print("Meilleurs paramètres:", best_params_ran)
print("Exactitude correspondante:", best_accuracy_ran)
```

```
Meilleurs paramètres: {'n_estimators': 10, 'max_depth': 7}
Exactitude correspondante: 0.9766081871345029
```

FIGURE IV.5 – implementation Random Forest

Nous avons appliqué l'algorithme des forêts aléatoires en utilisant l'implémentation RandomForestClassifier de scikit-learn. Afin d'optimiser ses performances, nous avons mené une recherche exhaustive par validation croisée pour déterminer les meilleurs hyperparamètres. Les meilleur hyperparamètres d'arbre est 10 et la profendeurs est de 7.

IV.4.5 Réseaux de neurones artificiels (ANN)

Inspirés du fonctionnement biologique des neurones, ces modèles connectionnistes peuvent apprendre des représentations profondes des données. C'est l'un des algorithmes les plus efficace avec des grandes quantités de données, leur principal défi reste la définition de leur architecture et l'optimisation des hyperparamètres.

```
model = Sequential()

model.add(Dense(units=30,activation='relu'))

model.add(Dense(units=15,activation='relu'))

model.add(Dense(units=1,activation='sigmoid'))

model.compile(loss='binary_crossentropy', optimizer='adam',metrics=['accuracy'])
```

```
early_stop = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=3)
```

```
model.fit(x=X_train,
          y=y_train,
          epochs=600,
          validation_data=(X_test, y_test), verbose=1,
          callbacks=[early_stop]
          )
```

FIGURE IV.6 – Implementation réseaux de neurones artificiels

Pour appliquer cet algorithme, nous avons utilisé keras un api de la bibliothèque tensorflow, ce réseau comporte une couche d'entrée, deux couches cachées entièrement connectées de 30 et 15 neurones avec une fonction d'activation ReLU, et une couche de sortie avec un unique neurone et une fonction d'activation sigmoïde pour la classification binaire.

IV.4.6 Régression logistique

Il s'agit d'un modèle linéaire statistique généralisé qui prédicte la probabilité d'appartenir à une classe en fonction des variables explicatives. Malgré son âge, il demeure une référence solide et compréhensible en termes de classification binaire.

```
log = LogisticRegression()
```

```
log.fit(X_train,y_train)
```

```
▸ LogisticRegression
```

```
y_pred_log = log.predict(X_test)
accuracy_log = accuracy_score(y_test,y_pred_log)
accuracy_log
```

FIGURE IV.7 – Implementation de régression logistique

Nous avons opté pour l'évaluation de ces algorithmes car ils englobent une variété d'approches (par modèle, linéaire, non linéaire, paramétrique, non paramétrique) et sont parmi

les plus couramment employés pour les problèmes de diagnostic médical binaire tels que la prédiction de cancer.

IV.5 Méthodologie expérimentale

Afin de comparer de manière juste et significative les performances des différents algorithmes de classification sur le problème de prédiction du cancer du sein, une méthodologie rigoureuse a été mise en place :

IV.5.1 Division et prétraitement des données

Nous avons divisé le jeux de donnée en deux sous-ensemble aléatoirement avec la méthode train_test_split de sklearn :
70% pour constituer l'ensemble d'entraînement et 30% réservés pour l'ensemble de test afin d'évaluer les performances finales.
Ensuite, nous avons mis à l'échelle les données pour que toutes les caractéristiques soient sur la même échelle.parce que certains algorithmes de machine learning sont sensibles aux différentes échelles des données. Pour cela, nous avons utilisé le MinMaxScaler, qui transforme les valeurs de chaque caractéristique pour qu'elles soient entre 0 et 1.

IV.5.2 Métriques d'évaluation

Pour évaluer les performances de chaque modèle sur les données de test, nous avons utilisées Plusieurs métriques standards tels que :

- L'exactitude (accuracy) représentant la proportion d'observations correctement classées.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- La précision (precision) évaluant la capacité du modèle à éviter les faux positifs.

$$\textbf{precision} = \frac{TP}{TP + FP}$$

- Le rappel (recall) mesurant sa capacité à détecter les cas positifs (cancers).

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

- La f1-score combinant de manière harmonique la précision et le rappel.

$$\textbf{F1-score} = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

✧ **TP :** Nombre d'images correctement classifiées comme patients atteints d'une tumeur.

✧ **FP :** Nombre d'images faussement classifiéees comme patients atteints d'une tumeur.

- ✧ **FN :** Nombre d'images faussement classifiéees comme patients non atteints d'une tumeur.
- ✧ **TN :** Nombre d'images correctement classifiées comme patients non atteints d'une tumeur.

Le choix de ces métriques complémentaires permet une évaluation complète, essentielle dans un contexte médical où il est crucial de minimiser les erreurs de classification.

IV.5.3 Résultats expérimentaux et analyse comparative des algorithmes

IV.5.3.1 Les résultats obtenus pour chaque algorithme (métriques de performance)

Voici les résultats que nous avons obtenues pour chaque algorithme :

IV.5.3.1.1 K-plus proches voisins (KNN) : En analysant les résultats de l'algorithme KNN, on peut voir que les scores sont bons. La précision et le rappel sont presque parfaits à 0,98 pour les deux classes. Le F1-score qui combine les deux est autour de 0,98. Même s'il y a un peu plus d'exemples dans une classe que dans l'autre, ça n'a pas eu trop d'impact car KNN a réussi à bien classer la plupart des exemples des deux côtés. En géneral, un accuracy de 0,98 est un excellent score. Du coup, on peut dire que KNN marche vraiment bien pour ce problème de classification binaire sur ces données.

	precision	recall	f1-score	support
0	0.98	0.97	0.97	59
1	0.98	0.99	0.99	112
accuracy			0.98	171
macro avg	0.98	0.98	0.98	171
weighted avg	0.98	0.98	0.98	171

FIGURE IV.8 – Rapport de classification KNN

IV.5.3.1.2 Arbre de décision : Pour les arbres de décision, les scores sont moins bon que que ceux du KNN. La précision est correcte avec 0,90 pour la classe 0 et excellente avec 0,95 pour la classe 1, tandis que le rappel est parfait avec 0,92 et 0,95 respectivement. Ainsi, les F1-scores, combinant précision et rappel, sont de 0,91 et 0,95, démontrant de bonnes performances globales avec une accuracy de 0,94, malgré un léger déséquilibre entre les classes.

	precision	recall	f1-score	support
0	0.90	0.92	0.91	59
1	0.95	0.95	0.95	112
accuracy			0.94	171
macro avg	0.93	0.93	0.93	171
weighted avg	0.94	0.94	0.94	171

FIGURE IV.9 – Rapport de classification arbre de décision

IV.5.3.1.3 Machines a vecteurs de support (SVM) : Pour l'algorithme SVM, les résultats sont excellents avec une précision de 1,00 pour la classe 0 et de 0,98 pour la classe 1, et des F1-scores de 0,98 et 0,99 respectivement. L'accuracy globale est de 0,99, démontrant une performance quasi parfaite malgré un déséquilibre entre les classes. Les moyennes des scores de précision, rappel et F1 sont toutes autour de 0,99, indiquant une performance équilibrée et robuste.

	precision	recall	f1-score	support
0	1.00	0.97	0.98	59
1	0.98	1.00	0.99	112
accuracy			0.99	171
macro avg	0.99	0.98	0.99	171
weighted avg	0.99	0.99	0.99	171

FIGURE IV.10 – Rapport de classification SVM

IV.5.3.1.4 Régression logistique : Pour l'algorithme de régression logistique, les résultats sont excellents avec une précision de 1,00 pour la classe 0 et de 0,98 pour la classe 1, et des F1-scores de 0,98 et 0,99 respectivement. L'accuracy globale est de 0,99, démontrant une performance presque parfaite malgré un déséquilibre entre les classes. Les moyennes des scores de précision, rappel et F1 sont toutes autour de 0,99, indiquant une performance équilibrée et fiable.

	precision	recall	f1-score	support
0	1.00	0.97	0.98	59
1	0.98	1.00	0.99	112
accuracy			0.99	171
macro avg	0.99	0.98	0.99	171
weighted avg	0.99	0.99	0.99	171

FIGURE IV.11 – Rapport de classification regression logistique

IV.5.3.1.5 Réseaux de neurones : Pour les réseaux de neurones, les résultats sont excellents avec une précision de 1,00 pour la classe 0 et de 0,99 pour la classe 1, et des F1-scores de 0,99 et 1,00 respectivement. L'accuracy globale est de 0,99, démontrant une performance quasi parfaite malgré un déséquilibre entre les classes. Les moyennes des scores de précision, rappel et F1 sont toutes autour de 0,99 à 1,00, indiquant une performance extrêmement fiable et équilibrée.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	59
1	0.99	1.00	1.00	112
accuracy			0.99	171
macro avg	1.00	0.99	0.99	171
weighted avg	0.99	0.99	0.99	171

FIGURE IV.12 – Rapport de classification ANN

IV.5.3.1.6 Random Forest : Pour les réseaux de neurones, les résultats sont excellents avec une précision de 1,00 pour la classe 0 et de 0,99 pour la classe 1, et des F1-scores de 0,99 et 1,00 respectivement. L'accuracy globale est de 0,99, démontrant une performance quasi parfaite malgré un déséquilibre entre les classes. Les moyennes des scores de précision, rappel et F1 sont toutes autour de 0,99 à 1,00, indiquant une performance extrêmement fiable et équilibrée.

Rapport de classification :

	precision	recall	f1-score	support
0	0.97	0.97	0.97	59
1	0.98	0.98	0.98	112
accuracy			0.98	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.98	0.98	0.98	171

FIGURE IV.13 – Rapport de classification Random Forest

IV.5.4 Analyse comparative

Les précisions obtenues pour les différents algorithmes sont les suivantes : ANN (0.9942), SVM (0.9883), Régression Logistique (0.9883), KNN (0.9825), Random Forest (0.9766), et Arbre de décision (0.9649). L'ANN présente la meilleure précision, suivi de près par le SVM et la Régression Logistique, qui ont des performances identiques. Le KNN et le Random Forest sont légèrement moins précis, tandis que l'Arbre de décision est le moins performant. Ainsi, pour une précision maximale et des performances optimales, l'ANN est le choix idéal.

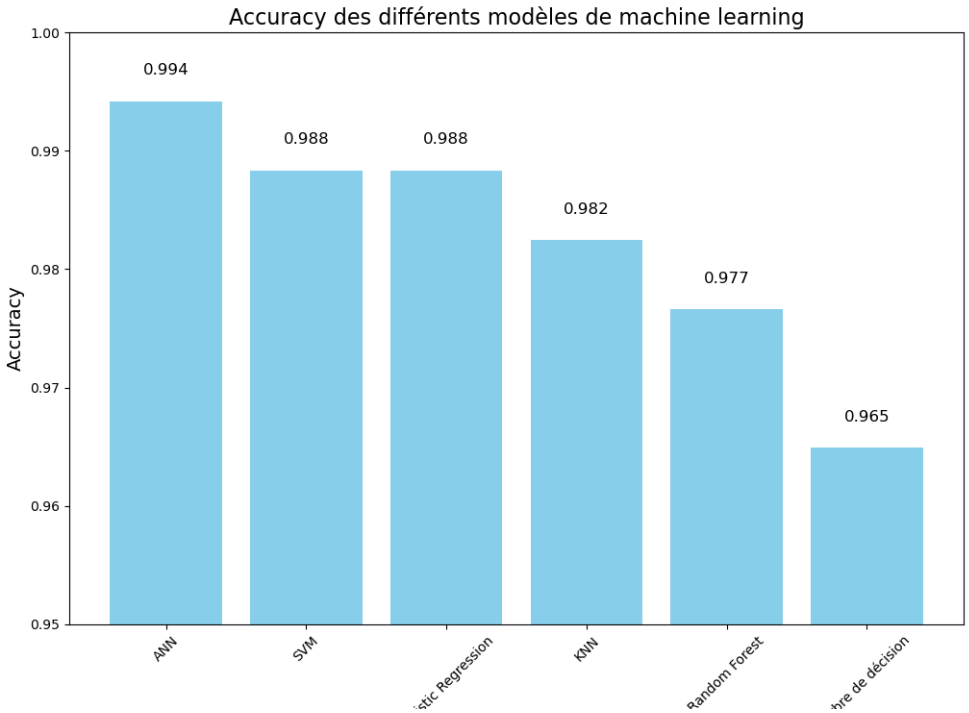


FIGURE IV.14 – Analyse comparative des précisions des algorithmes

Partie 2 : diagnostic et classification

IV.6 Le choix du dataset

Dans le cadre de la réalisation de notre application de diagnostic et classification du cancer du sein, nous avons opté pour le dataset Breast Histopathology Images, qui contient des images d'histopathologie de spécimens de cancer du sein. Ce dataset comprend 277,524 patchs de taille 50x50 pixels, extraits de 162 images de lames entières scannées à une résolution de 40x. Parmi ces patchs, 78,786 sont négatifs pour le carcinome canalaire infiltrant (IDC) et 78,786 sont positifs pour l'IDC.[46]
L'objectif principal de ce dataset est de faciliter le développement de méthodes automatiques pour délimiter et classifier les régions contenant de l'IDC dans les lames de biopsie. Cela aide les pathologistes à évaluer plus précisément et rapidement l'agressivité du cancer du sein, une tâche cruciale pour le diagnostic et le traitement clinique.

IV.7 Choix du modéles

Pour le choix des modéles, Nous avons utilisé des réseaux de neurones convolutifs (CNN).Ils sont bien adaptés pour la classification d'images parce qu'ils peuvent détecter des caractéristiques locales dans les images en apprenant des filtres. Ces filtres, ou noyaux, permettent au réseau de détecter des motifs dans l'image d'entrée, ce qui aide à identifier des caractéristiques complexes.

En plus de concevoir nos propres réseaux, nous avons également exploité plusieurs modèles préentraînés qui ont une bonne réputation dans la reconnaissance d'images. Parmi ceux-ci, nous avons utilisé DenseNet201, VGG19 et Resnet50. Ces modèles préentraînés ont l'avantage de disposer de grandes quantités de données sur lesquelles ils ont été initialement formés, ce qui leur permet de transmettre des connaissances pertinentes à notre propre travail de classification d'image mamographique.

Pour l'implémentation de nos modèle, nous avons utilisé le frameworks de deep learning TensorFlow avec son api keras.

IV.7.1 Modéle CNN

Lors de la conception de notre propre CNN, nous avons soigneusement sélectionné et optimisé les paramètres du modèle a fin de trouver un meilleur résultat possible.
Le Modéle contient 4 couche de convolution 2D qui ont les configurations suivantes :

- ✦ La première couche avec un filtre de 16 de taille (3,3) avec une fonction d'activation ReLu, cette couche prend en entrée des images de taille (50, 50, 3).
- ✦ La deuxième couche avec un filtre de 32 de taille (3,3) avec ReLu.
- ✦ La troisième couche avec un filtre de 64 de taille (3,3) avec Relu.
- ✦ La Quatrième couche avec un filtre de 64 de taille (3,3) avec Relu.

Pour chaque couche nous avons utilisés des couche maxpooling de taille (2,2) pour réduire la taille et les dimensions des données, tout en conservant les caractéristiques importantes. Ensuite nous avons ajouter une couche Flatten pour platir nos données en un vecteur 1D et Dropout pour désactiver aléatoirement 50% des neurones pendant l'entraînement pour réduire le surapprentissage.

Pour conclure, deux couches denses ont été intégrées :

- ✦ Une couche de 256 unités avec ReLU.
- ✦ Une couche de 1 unité (couche de sortie) avec une fonction d'activation sigmoïde.

Voici le résumé des paramètres :

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 48, 48, 16)        448

 max_pooling2d (MaxPooling2  (None, 24, 24, 16)        0
 D)

 conv2d_1 (Conv2D)           (None, 22, 22, 32)        4640

 max_pooling2d_1 (MaxPoolin  (None, 11, 11, 32)        0
 g2D)

 conv2d_2 (Conv2D)           (None, 9, 9, 64)          18496

 max_pooling2d_2 (MaxPoolin  (None, 4, 4, 64)          0
 g2D)

 conv2d_3 (Conv2D)           (None, 2, 2, 64)          36928

 max_pooling2d_3 (MaxPoolin  (None, 1, 1, 64)          0
 g2D)

 flatten (Flatten)           (None, 64)                0

 dense (Dense)               (None, 256)               16640

 activation (Activation)     (None, 256)               0

 dropout (Dropout)           (None, 256)               0

 dense_1 (Dense)             (None, 1)                 257

 activation_1 (Activation)   (None, 1)                 0

=================================================================
Total params: 77409 (302.38 KB)
Trainable params: 77409 (302.38 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

IV.7.2 Modéle pré-entrainé

IV.7.2.1 VGG19

VGG19 est réseau de neurones convolutionnels proposés par K. Simonyan et A. Zisserman de l'université d'Oxford et qui a acquis une notoriété en gagnant la compétition ILSVRC (ImageNet Large Scale Visual Recognition Challenge) en 2014. Le modèle a atteint une précision de 92.7% sur Imagenet ce qui est un des meilleurs scores obtenus., il est connu pour son architecture et sa profondeur simplifiée avec des couches convolutives de petite taille (3x3).Nous avons utilisé cette architecture comme base pour notre modèle de classification d'images, qui nous permette d'en tirer parti de ses capacités d'extraction de caractéristiques puissantes.[47]

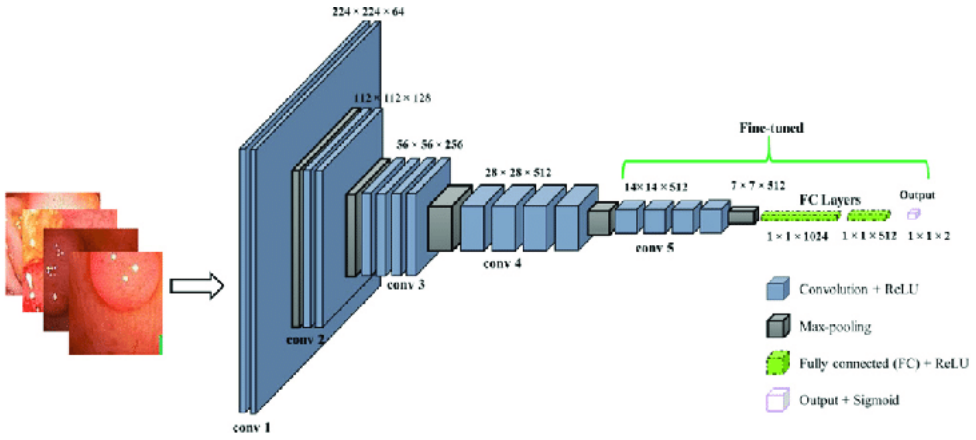


FIGURE IV.15 – Modéle vgg19

IV.7.2.2 Densnet201

Densnet201 est une architecture de réseau neuronal convolutif profond (CNN) développée par Huang et al. en 2017[48]. Elle est basée sur le concept de connectivité dense, où chaque couche est connectée à toutes les couches précédentes, ce qui permet une meilleure propagation des informations à travers le réseau.

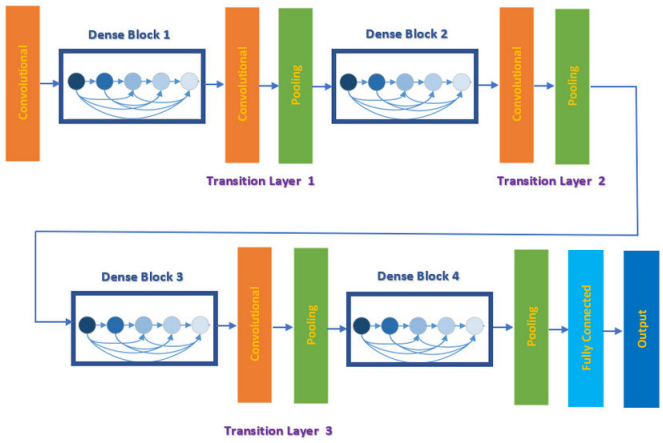


FIGURE IV.16 – Modéle Densnet201

IV.7.2.3 Resnet50

ResNet-50 est un réseau de neurones à convolution d'une profondeur de 50 couches proposée par Kaiming He et al. en 2015, le réseau de neurones a appris des représentations avec de nombreuses caractéristiques pour une grande variété d'images. [49].

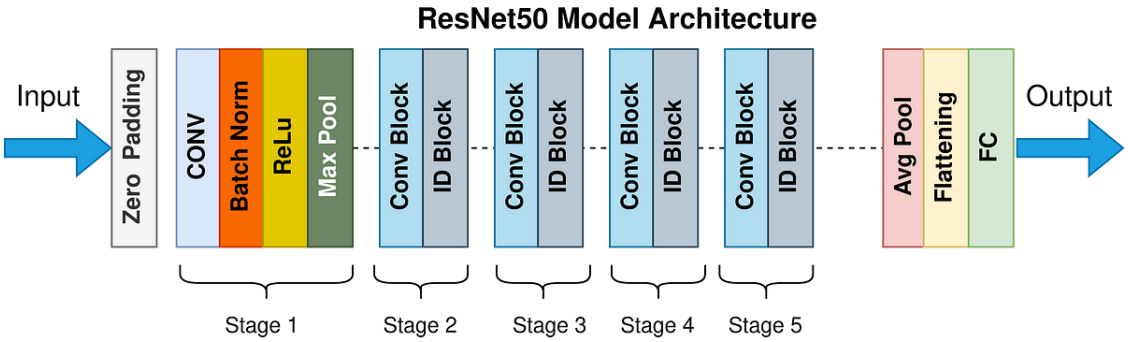


FIGURE IV.17 – Modéle Resnet50

IV.7.2.4 Architecture des modéle pré-entrainé

Pour chaque modéle, nous les avons chargés avec Keras comme base de notre réseau. Ensuite, Nous avons ajouté des couches supplémentaires au-dessus des modèle pour adapter l'architecture à notre tâche de classification binaire.

- **Flatten** :Platir nos données en un vecteur 1D.
- **Dropout (0.5)** : Désactive aléatoirement 50% des neurones pendant l'entraînement pour réduire le surapprentissage.
- **Dense (1 unité, activation 'sigmoid')** : Fournit la sortie finale pour la classification binaire (0 ou 1).

IV.7.2.4.1 Résumé des paramètres VGG19

```
Model: "sequential_16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg19 (Functional)           (None, 1, 1, 512)         20024384
flatten_16 (Flatten)         (None, 512)               0
dropout_42 (Dropout)         (None, 512)               0
dense_56 (Dense)             (None, 1)                 513
Total params: 20024897 (76.39 MB)
Trainable params: 20024897 (76.39 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

IV.7.2.4.2 Résumé des paramètres Densnet201

```
Model: "sequential_17"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 densenet201 (Functional)    (None, 1, 1, 1920)        18321984

 flatten_17 (Flatten)        (None, 1920)              0

 dropout_43 (Dropout)        (None, 1920)              0

 dense_57 (Dense)            (None, 1)                 1921

=================================================================
Total params: 18323905 (69.90 MB)
Trainable params: 18094849 (69.03 MB)
Non-trainable params: 229056 (894.75 KB)
_________________________________________________________________
```

IV.7.2.4.3 Résumé des paramètres Resnet50

```
Model: "sequential_19"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resnet50 (Functional)       (None, 2, 2, 2048)        23587712

 flatten_19 (Flatten)        (None, 8192)              0

 dropout_45 (Dropout)        (None, 8192)              0

 dense_59 (Dense)            (None, 1)                 8193

=================================================================
Total params: 23595905 (90.01 MB)
Trainable params: 23542785 (89.81 MB)
Non-trainable params: 53120 (207.50 KB)
_________________________________________________________________
```

IV.8 Préparation des données

IV.8.1 Chargement du dataset a partir de Kaggle

Pour télécharger le dataset directement depuis kaggle, nous avons utiliser la bibliothèque Kaggle API qui nous permet d'interagir avec la plateforme avec une ligne de commande pour télécharger des datasets.

```
!kaggle datasets download alaminbhuyan/breast-histopathology-images

Dataset URL: https://www.kaggle.com/datasets/alaminbhuyan/breast-histopathology-images
License(s): unknown
Downloading breast-histopathology-images.zip to /content
100% 926M/929M [00:58<00:00, 20.5MB/s]
100% 929M/929M [00:58<00:00, 16.6MB/s]
```

FIGURE IV.18 – Code chargement du dataset

Dans ce code nous avons utiliser la commande !kaggle pour importer le dataset depuis un lien.

```
import zipfile
import os


zip_file_path = "/content/breast-histopathology-images.zip"
extract_path = "/content/breast-histopathology-images"

if not os.path.exists(extract_path):
    os.makedirs(extract_path)


with zipfile.ZipFile(zip_file_path, 'r') as zip_ref:
    zip_ref.extractall(extract_path)

extracted_files = os.listdir(extract_path)
extracted_files
```

```
['IDC_regular_ps50_idx5']
```

FIGURE IV.19 – Code d'extraction

Ensuite nous avons extrait le dataset car il été sous format zip avec la bibliothéque zipfile.

IV.8.2 Charger les chemins

Pour chaque classe nous avons chargé son chemin d'accés respectif.

```
my_data_dir = '/content/breast-histopathology-images/IDC_regular_ps50_idx5'
```

```
os.listdir(my_data_dir)
```

```
['negative_IDC', 'positive_IDC']
```

```
negative_path = my_data_dir+'/negative_IDC/'
positive_path = my_data_dir+'/positive_IDC/'
```

FIGURE IV.20 – code chargement chemins d'accés

IV.8.3 Division des données

Nous avons divisé nos images en ensembles d'entraînement, de validation et de test avec la fonction train_test_split de sklearn.

```
IDC_négatifs_train, IDC_négatifs_val_test = train_test_split(IDC_négatifs_paths, test_size=0.2, random_state=104)
IDC_négatifs_val, IDC_négatifs_test = train_test_split(IDC_négatifs_val_test, test_size=0.5, random_state=104)

IDC_positifs_train, IDC_positifs_val_test = train_test_split(IDC_positifs_paths, test_size=0.2, random_state=104)
IDC_positifs_val, IDC_positifs_test = train_test_split(IDC_positifs_val_test, test_size=0.5, random_state=104)
```

```
print("Nombre d'images IDC négatives :")
print("Train:", len(IDC_négatifs_train), "Validation:", len(IDC_négatifs_val), "Test:", len(IDC_négatifs_test))

print("\nNombre d'images IDC positives :")
print("Train:", len(IDC_positifs_train), "Validation:", len(IDC_positifs_val), "Test:", len(IDC_positifs_test))
```

```
Nombre d'images IDC négatives :
Train: 63028 Validation: 7879 Test: 7879

Nombre d'images IDC positives :
Train: 63028 Validation: 7879 Test: 7879
```

FIGURE IV.21 – code divison des données

80% pour l'entraînement et 10% pour la validation et 10% pour le test.

IV.8.4 Chargement et transformation des images

Aprés avoir divisé nos données, nous avons chargé les images avec leur étiquette depuis leurs ensembles respectives (train, test, validation) en les transformants en tableaux numpy avec une taille de (50,50) pixels.

```
from tensorflow.keras.preprocessing.image import load_img, img_to_array
import numpy as np

def load_images(file_paths, label):
    images = []
    labels = []
    for path in file_paths:
        img = load_img(path, target_size=(50, 50))
        img_array = img_to_array(img)
        images.append(img_array)
        labels.append(label)
    return np.array(images), np.array(labels)


X_train_négatifs, y_train_négatifs = load_images(IDC_négatifs_train, 0)
X_train_positifs, y_train_positifs = load_images(IDC_positifs_train, 1)

X_val_négatifs, y_val_négatifs = load_images(IDC_négatifs_val, 0)
X_val_positifs, y_val_positifs = load_images(IDC_positifs_val, 1)

X_test_négatifs, y_test_négatifs = load_images(IDC_négatifs_test, 0)
X_test_positifs, y_test_positifs = load_images(IDC_positifs_test, 1)


X_train = np.concatenate([X_train_négatifs, X_train_positifs])
y_train = np.concatenate([y_train_négatifs, y_train_positifs])

X_val = np.concatenate([X_val_négatifs, X_val_positifs])
y_val = np.concatenate([y_val_négatifs, y_val_positifs])

X_test = np.concatenate([X_test_négatifs, X_test_positifs])
y_test = np.concatenate([y_test_négatifs, y_test_positifs])

print("Ensemble d'entraînement :", X_train.shape, y_train.shape)
print("Ensemble de validation :", X_val.shape, y_val.shape)
print("Ensemble de test :", X_test.shape, y_test.shape)
```

```
Ensemble d'entraînement : (126056, 50, 50, 3) (126056,)
Ensemble de validation : (15758, 50, 50, 3) (15758,)
Ensemble de test : (15758, 50, 50, 3) (15758,)
```

FIGURE IV.22 – code chargement et transformation

IV.8.5 Augmentation de données

Après avoir chargé, divisé et transformé les images, nous arrivons à l'étape de l'augmentation des données. Cette étape est essentielle pour augmenter la taille et la diversité du dataset d'entraînement en générant de nouvelles données artificielles à partir des données existantes. Cette technique vise à améliorer la capacité du modèle à généraliser et à réduire le surapprentissage.

```
from tensorflow.keras.preprocessing.image import ImageDataGenerator

datagen = ImageDataGenerator(rotation_range=20, # Faire pivoter l'image de 20 degrés
                             width_shift_range=0.10, # Modifier la largeur de la photo de 10% maximum
                             height_shift_range=0.10, # Modifier la hauteur de la photo de 10% maximum
                             shear_range=0.1, # shear signifie couper une partie de l'image (max 10%)
                             zoom_range=0.1, # Zoom de 10% maximum
                             horizontal_flip=True, # Autorise le basculement horizontal
                             fill_mode='nearest' # Remplir les pixels manquants avec la valeur remplie la plus proche
                            )
train_generator = datagen.flow(X_train, y_train, batch_size=64)
```

FIGURE IV.23 – code augmentation d'image

Ce code charge des images, applique des augmentations de données (rotation, zoom, flip horizontal...),ces transforCes transformations préservent la sémantique des données tout en introduisant des variations, ce qui permet au modèle de mieux capturer les différentes perspectives.
Voici un exemple de l'augmentation de données appliquée sur une image

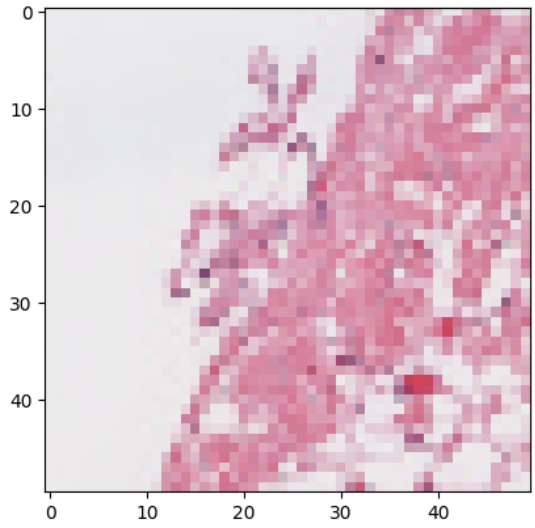


FIGURE IV.24 – Image original

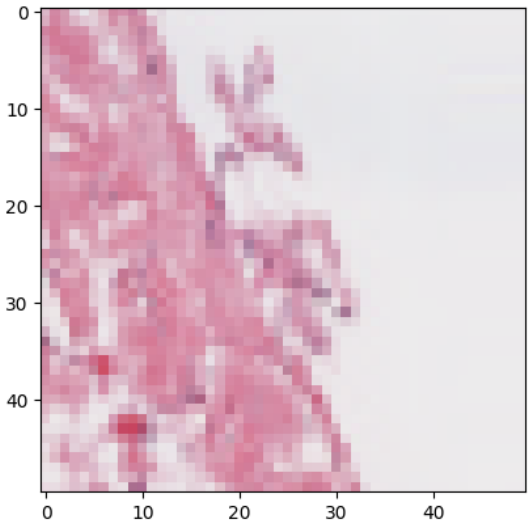


FIGURE IV.25 – Image augmenté

IV.9 Compilation et entraînement

Enfin, nous arrivons a la dernière étape de notre modéle. Cette étape est cruciale car elle transforme notre modèle théorique en un modèle fonctionnel qui peut faire des prédictions sur des données nouvelles.

```
model_vgg19.compile(optimizer=Adam(learning_rate=0.0001), loss='binary_crossentropy', metrics=['accuracy'])
early_stop = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)
history_vgg19 = model_vgg19.fit(train_generator, epochs=100, batch_size=64, validation_data=(X_val, y_val),callbacks=[early_stop])
```

FIGURE IV.26 – code compilation

Nous avons configuré notre modèle en utilisant l'optimiseur Adam avec un taux d'apprentissage de 0.0001, la fonction de perte binary_crossentropy pour les problèmes de classification binaire, et nous avons surveillé la métrique de précision (accuracy) pour évaluer les performances pendant l'entraînement.

IV.10 Résultats et discussion

Dans cette section, nous présentons les performances de notre modèle CNN ansi les autres architectures des CNN en termes d'accuracy, de loss et d'autres métriques d'évaluation sur l'ensemble de test, notamment le classification report et la matrice de confusion.

IV.10.1 Performances des modèles CNN

IV.10.1.1 Modéle CNN

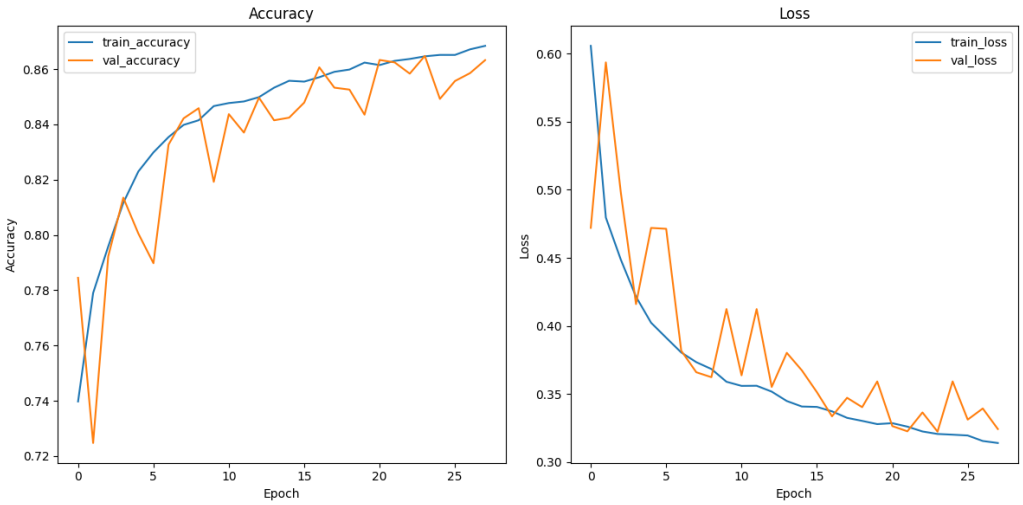


FIGURE IV.27 – accuracy et loss CNN

IV.10.1.2 Modéle VGG19

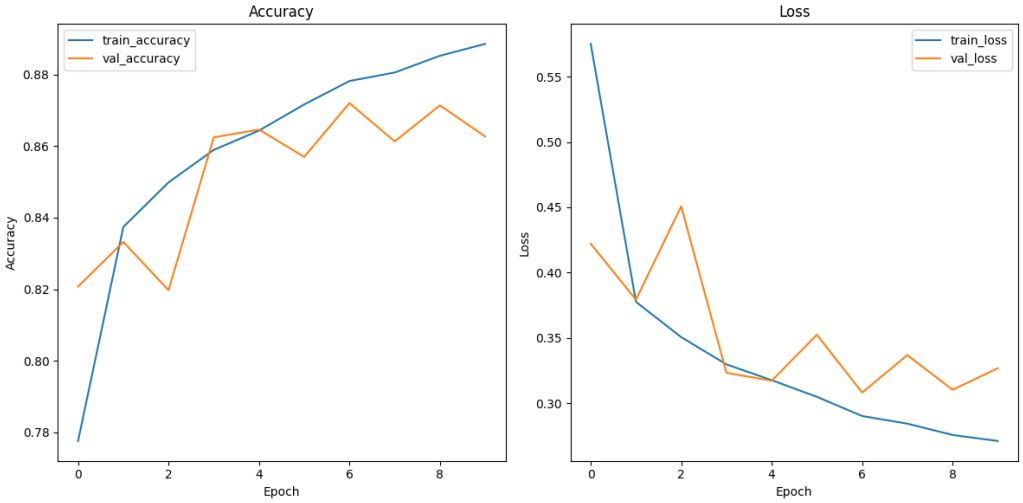


FIGURE IV.28 – accuracy et loss vgg19

IV.10.1.3 Modéle Densnet201

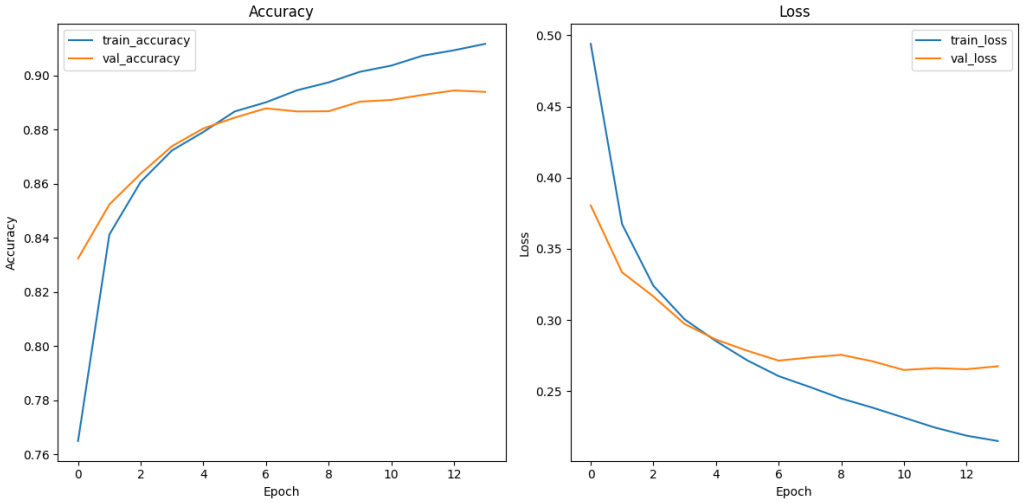


FIGURE IV.29 – accuracy et loss densnet201

IV.10.1.4 Modéle Resnet50

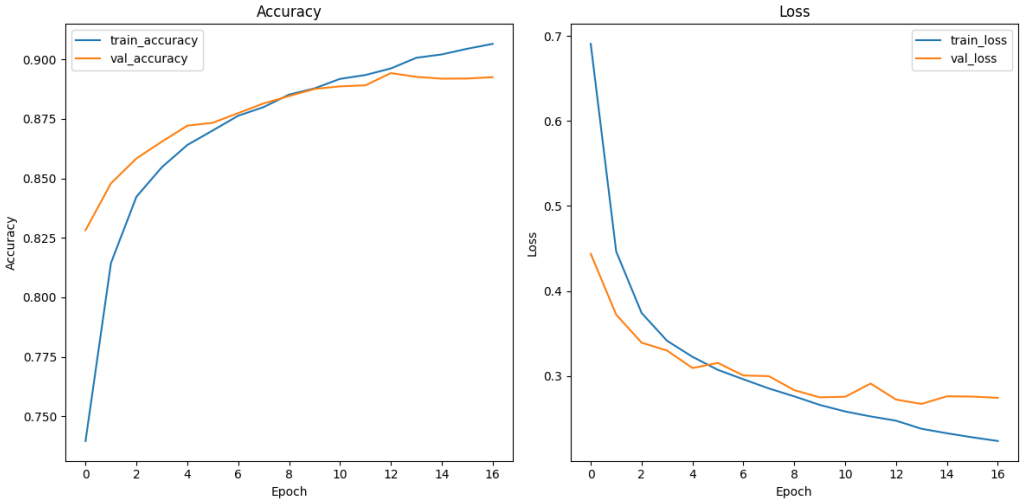


FIGURE IV.30 – accuracy et loss resnet50

Les figures ci-dessus montrent l'évolution de l'accuracy et de la loss pendant l'entraînement des modéles que nous avons utilisé. On constate une augmentation de l'accuracy et une diminution de la loss au fur et à mesure des epochs, ce qui indique que notre modèle apprend efficacement à partir des données d'entraînement.

IV.10.2 Évaluation sur l'ensemble de test

IV.10.2.1 Modéle CNN

IV.10.2.1.1 Classification repport

```
493/493 [==============================] - 1s 2ms/step
              precision    recall  f1-score   support

    négative       0.84      0.90      0.87      7879
    positive       0.89      0.83      0.86      7879

    accuracy                           0.86     15758
   macro avg       0.87      0.86      0.86     15758
weighted avg       0.87      0.86      0.86     15758
```

Les résultats montrent que le modèle CNN a une précision, un recall et un F1-score de 0.86 pour les classes "négative" et "positive", avec une accuracy globale de 0.86.

IV.10.2.1.2 Matrice de confusion

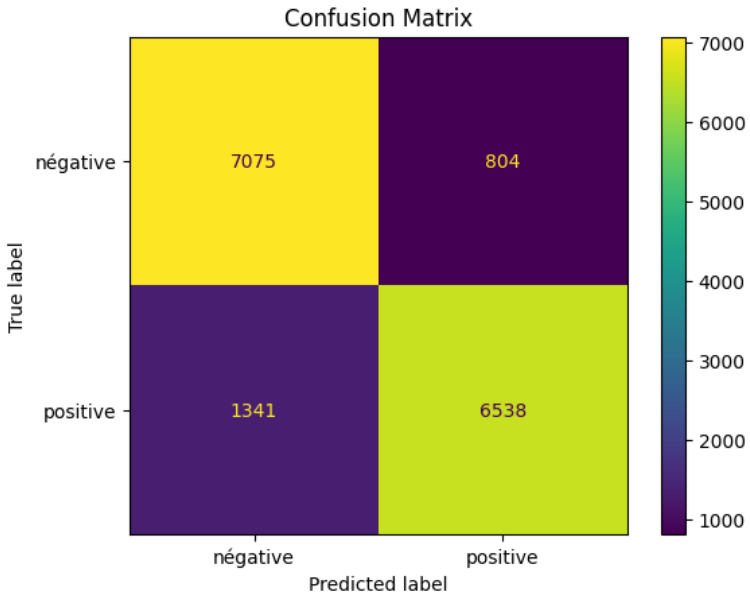


FIGURE IV.31 – Matrice de confusion

- **négative (Vrai) - négative (Prédiction) :** 7075 (vrais négatifs) - Le modèle a correctement prédit 7075 images négative.
- **négative (Vrai) - positive (Prédiction) :** 804 (faux négatifs) - Le modèle a incorrectement prédit 804 images négative comme positive.
- **positive (Vrai) - négative (Prédiction) :** 1341 (faux positifs) - Le modèle a incorrectement prédit 1341 images postive comme négative.
- **positive (Vrai) - positive (Prédiction) :** 6538 (vrais positive) - Le modèle a correctement prédit 6538 images positive.

IV.10.2.2 Modèle vgg19

IV.10.2.2.1 Classification repport

```
493/493 [==============================] - 8s 15ms/step
              precision    recall  f1-score   support

    négative       0.91      0.83      0.87      7879
    positive       0.84      0.91      0.88      7879

    accuracy                           0.87     15758
   macro avg       0.88      0.87      0.87     15758
weighted avg       0.88      0.87      0.87     15758
```

Le modèle vgg19 a montré que il a amélioré ses performances par rapport au modèle précédent.

IV.10.2.2.2 Matrice de confusion

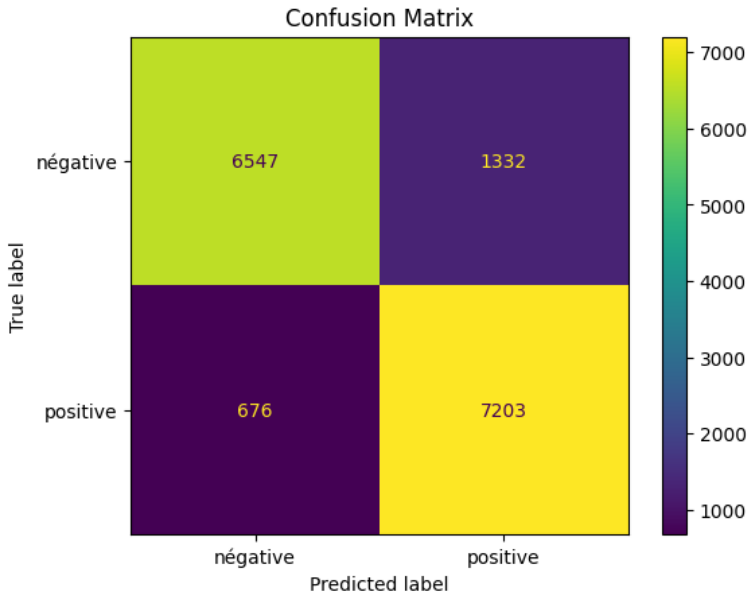


FIGURE IV.32 – Matrice de confusion vgg19

- **négative (Vrai) - négative (Prédiction) :** 6547 (vrais négatifs) - Le modèle a correctement prédit 6547 images négative.
- **négative (Vrai) - positive (Prédiction) :** 1332 (faux négatifs) - Le modèle a incorrectement prédit 1332 images négative comme positive.
- **positive (Vrai) - négative (Prédiction) :** 676 (faux positifs) - Le modèle a incorrectement prédit 676 images postive comme négative.
- **positive (Vrai) - positive (Prédiction) :** 7203 (vrais positifs) - Le modèle a correctement prédit 7203 images positive.

IV.10.2.3 Modéle Densnet201

IV.10.2.3.1 Classification repport

```
493/493 [==============================] - 13s 21ms/step
              precision    recall  f1-score   support

    négative       0.88      0.90      0.89      7879
    positive       0.90      0.88      0.89      7879

    accuracy                           0.89     15758
   macro avg       0.89      0.89      0.89     15758
weighted avg       0.89      0.89      0.89     15758
```

Le modèle densnet201 est plus performant que vgg19 et le modèle cnn que nous avons fait.

IV.10.2.3.2 Matrice de confusion

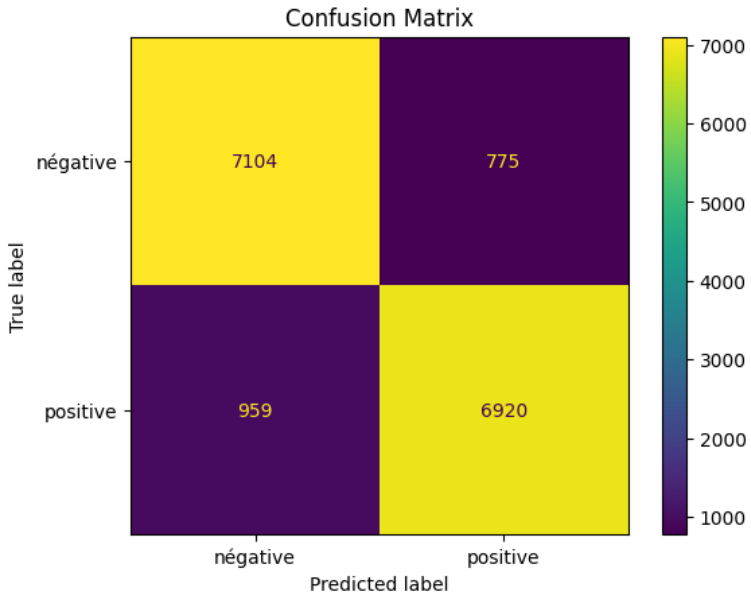


FIGURE IV.33 – Matrice de confusion densnet201

- **négative (Vrai) - négative (Prédiction) :** 7104 (vrais négatifs) - Le modèle a correctement prédit 7104 images négative.
- **négative (Vrai) - positive (Prédiction) :** 775 (faux négatifs) - Le modèle a incorrectement prédit 775 images négative comme positive.
- **positive (Vrai) - négative (Prédiction) :** 959 (faux positifs) - Le modèle a incorrectement prédit 959 images postive comme négative.
- **positive (Vrai) - positive (Prédiction) :** 6920 (vrais positifs) - Le modèle a correctement prédit 6920 images positive.

IV.10.2.4 Modèle Resnet50

IV.10.2.4.1 Classification repport

```
493/493 [==============================] - 7s 12ms/step
              precision    recall  f1-score   support

    négative       0.89      0.90      0.89      7879
    positive       0.90      0.89      0.89      7879

    accuracy                           0.89     15758
   macro avg       0.89      0.89      0.89     15758
weighted avg       0.89      0.89      0.89     15758
```

Le modèle resnet50 est plus performant que tous les modèles précédents.

IV.10.2.4.2 Matrice de confusion

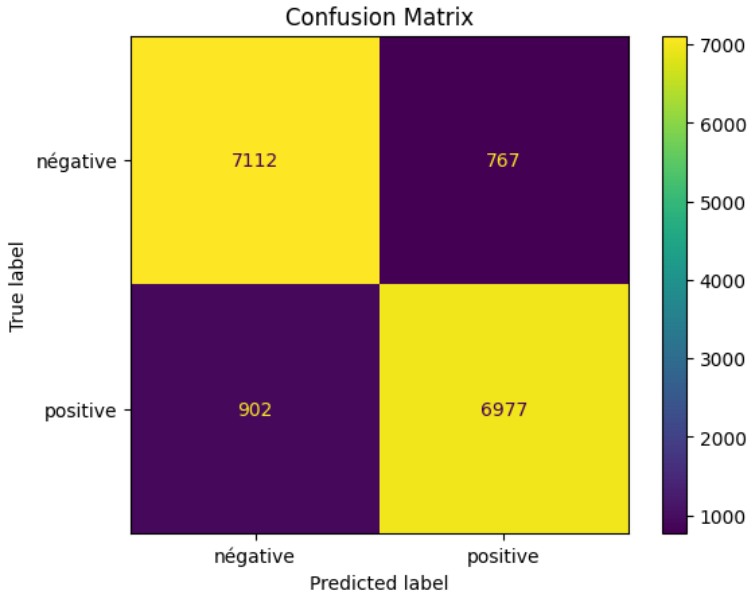


FIGURE IV.34 – Matrice de confusion resnet50

- **négative (Vrai) - négative (Prédiction) :** 7112 (vrais négatifs) - Le modèle a correctement prédit 7112 images négative.
- **négative (Vrai) - positive (Prédiction) :** 767 (faux négatifs) - Le modèle a incorrectement prédit 767 images négative comme positive.
- **positive (Vrai) - négative (Prédiction) :** 902 (faux positifs) - Le modèle a incorrectement prédit 902 images postive comme négative.
- **positive (Vrai) - positive (Prédiction) :** 6977 (vrais positifs) - Le modèle a correctement prédit 6977 images positive.

IV.10.3 Comparaison des modèles

IV.10.3.1 Précision des modèles

On a fait cette comparaison en termes de précision du modèle, temps d'exécution entre les 04 modèles.

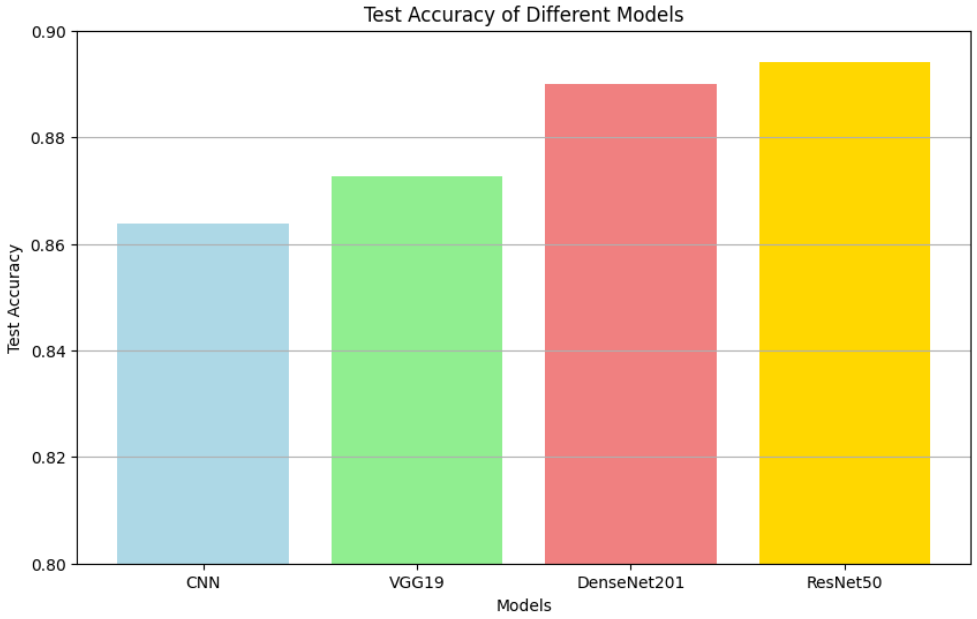


FIGURE IV.35 – Comparaison des modèles

Nous remarquons que la précision du modéle ResNet50 est supérieur aux autres modèles.

IV.10.3.2 Temps d'exécution des modèles

```
1- CNN :  493/493 [==============================] - 1s 2ms/step

2- ResNet50 : 493/493 [==============================] - 7s 12ms/step

3- VGG19 : 493/493 [==============================] - 8s 15ms/step

4- Densnet201: 493/493 [==============================] - 13s 21ms/step
```

IV.11 Conclusion

Au cours de ce chapitre, nous avons examiné et comparé divers algorithmes d'apprentissage supervisé afin de prédire le cancer du sein en utilisant des données tabulaires. L'ANN a été le meilleur algorithme évalué parmi les algorithmes - KNN, arbre de décision, SVM, régression logistique, réseaux de neurones (ANN) et forêt aléatoire - avec une précision de 0,9942. Les résultats du SVM et de la régression logistique ont également été remarquables, avec une précision de 0,9883. Malgré une légère inexactitude, le KNN, la forêt aléatoire et l'arbre de décision demeurent des alternatives réalisables. Donc, afin de maximiser la précision de ces données, l'utilisation de l'ANN semble être la solution idéale.

En ce qui concerne le diagnostic sur base d'imagerie médicale, nous avons créé un réseau de neurones convolutif (CNN) ainsi que trois modèles pré-entraînés : VGG19, DenseNet201 ainsi que ResNet50. La précision du modèle CNN a été de 0,86. Le modèle ResNet50 a démontré sa performance la plus élevée, dépassant les autres modèles. Le DenseNet201 a démontré des résultats inférieurs à ceux de ResNet50, mais plus performants que ceux de VGG19 et du CNN. VGG19 présentait une performance inférieure à celle de ResNet50 et DenseNet201, mais était plus précis que le CNN.

En conclusion, afin d'optimiser la classification des images médicales, le modèle ResNet50 pré-entraîné est le meilleur choix parmi les différentes approches évaluées. Elle est supérieure aux performances du CNN personnalisé et des modèles DenseNet201 et VGG19. Ces résultats encourageants ouvrent la voie à d'autres améliorations et éventuels usages cliniques.

ChapitreV

Réalisation d'une application de déploiement

V.1 Introduction

Dans ce chapitre, nous proposons a développer un outil d'aide à la décision basé sur les modèle que nous avons déjà créé pour la prédiction et la classification des tumeurs mammaires. Notre objectif est de concevoir une application web qui permet au professionnel de la santé de l'utiliser dans leur pratique clinique.

Nous commencerons par présenter la bibliothèque Taipy et expliquer son choix pour ce projet, ensuite nous présenterons notre application en fournissant les détails de son fonctionnement.

V.2 Le choix de la bibliothèque Taipy

Taipy est une bibliothèque open source Python conçue pour faciliter le développement d'applications Web basées sur les données.
Taipy couvre à la fois le front-end et le back-end. Il a été conçu pour accélérer le développement d'applications, depuis les prototypes initiaux jusqu'aux applications prêtes pour la production.[51]
Nous avons choisi taipy car elle permet de concevoir des applications facilement , elle offre la possibilité d'utiliser le markdown pour la partie front ce qui rend la tache vraiment sim.

V.3 Les interfaces

Les interfaces de notre application se composent de deux menus : le premier est dédié à la partie diagnostic, et le second à la partie prédiction.

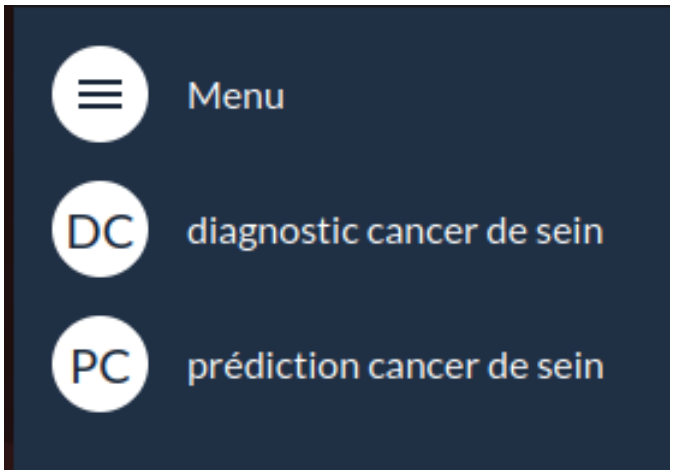


Figure V.1 – Menu interface

V.3.1 Menu diagnostic

Le menu diagnostic permet à l'utilisateur de charger des images a diagnostiquer, d'analyser ces images à l'aide de notre modèle de classification d'images , et de recevoir des résultats instantanés via des notifications.

FIGURE V.2 – Menu diagnostic

V.3.2 Menu prédiction

Le menu prédiction offre à l'utilisateur d'introduire les données sur la tumeur, d'analyser ces données à l'aide de notre modèle de prédiction que nous avons déjà conçus , et de recevoir des résultats instantanés via des notifications.

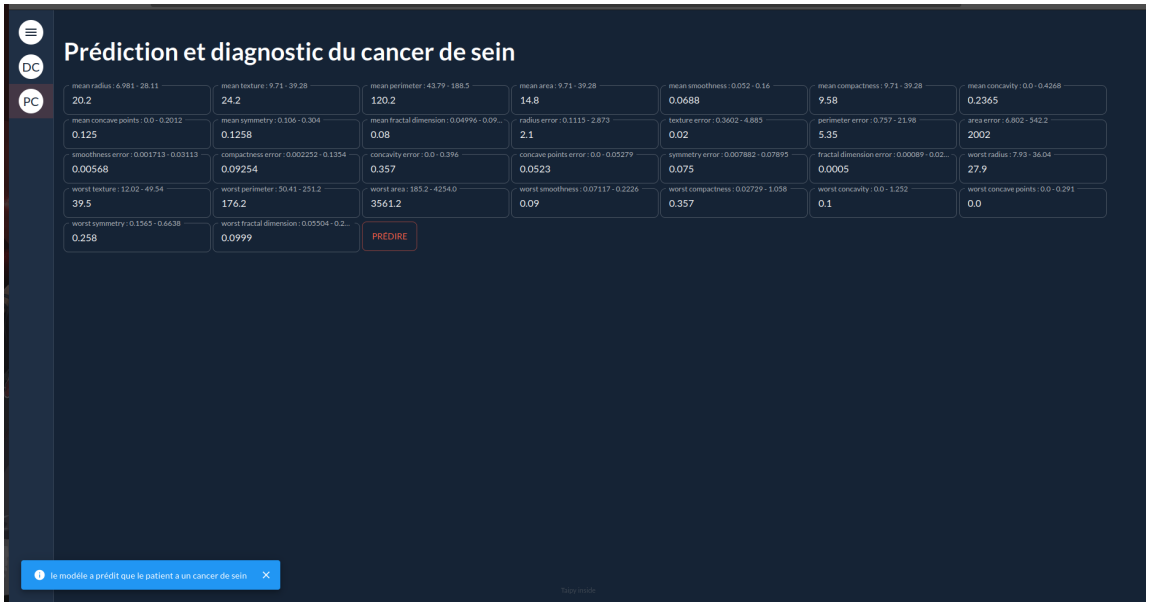


FIGURE V.3 – Menu prédiction

V.4 conclusion

Au cours de ce chapitre, nous avons créé un outil de prise de décision qui repose sur les modèles de prédiction et de classification des tumeurs mammaires. Grâce à l'utilisation de la bibliothèque Taipy, nous avons développé une application web intuitive et interactive destinée pour les professionnels de la santé. . Grâce à ses menus de prédiction et de diagnostic, cette application offre la possibilité de charger des données, de réaliser des prédictions et de visualiser les résultats de manière efficace. En somme, notre outil simplifie la prise de décisions cliniques en se basant sur des analyses précises et fiables.

Conclusion générale

Ce mémoire a porté sur l'application des techniques d'IA pour le diagnostic et la prédiction du cancer du sein, en se concentrant sur l'utilisation de modèles de machine learning et de deep learning. L'objectif était ; dans un premier temps, de développer un modèle d'IA capable de prédire le cancer du sein sur la base d'un dataset comportant des caractéristiques de types textuelles et dans un second temps, de développer un modèle permettant la classification d'images histopathologiques afin de distinguer les tissus cancéreux des tissus sains.

Nous avons d'abord introduit les concepts fondamentaux liés au cancer du sein, en décrivant les différents types de tumeurs, les méthodes de dépistage, et les techniques de diagnostic actuelles. Cette section a mis en lumière les défis auxquels sont confrontés les cliniciens dans le diagnostic précoce du cancer du sein.

Ensuite, nous avons exploré les principes de l'apprentissage automatique et du deep learning, en soulignant leur pertinence dans le traitement des images médicales. Nous avons discuté des architectures de réseaux de neurones convolutifs **(CNN)** et de leur capacité à extraire des caractéristiques complexes des images.

La partie expérimentale de ce mémoire a impliqué l'utilisation du dataset Breast Cancer Winsconsin pour la partie prédiction et du dataset Breast Histopathology Images pour la partie classification/diagnostic du cancer du sein. Nous avons ainsi appliqué plusieurs méthodes comme KNN, ANN, SVM, les arbres de décision et la régression logistique pour la partie prédiction. Concernant la partie diagnostic, nous avons eu à traiter des images et à une problématique de classification et c'est la raison pour laquelle nous avons appliqué plusieurs modèles de réseau de neurones convolutifs pour classifier les images en tissus sains et cancéreux. Les performances des modèles ont été évaluées en termes de précision, de rappel, et de F1-score, démontrant un fort potentiel pour une aide à la décision.

En plus des expériences menées, nous avons également développé une application interactive utilisant la bibliothèque **Taipy**. Cette application permet aux cliniciens de télécharger des images et d'obtenir des prédictions en temps réel, facilitant ainsi l'intégration des modèles d'intelligence artificielle dans la pratique clinique.

Les résultats de nos expériences montrent que les modèles de deep learning, en particulier les CNN, peuvent aider considérablement au diagnostic du cancer du sein. L'application développée offre un outil puissant pour les professionnels de santé, leur permettant de prendre des décisions éclairées basées sur les analyses automatisées des images histopathologiques.

En conclusion, ce travail illustre le potentiel des technologies du machine learning pour transformer le diagnostic médical, dans le domaine de l'oncologie. Les avancées réalisées ouvrent la voie à des recherches futures et à des améliorations continues des modèles, visant à fournir des outils plus robustes et précis pour le diagnostic et la prédiction.

Bibliographie

[1] Organisation Mondiale de la santé
https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer.
Consulté le 18 janvier 2024

[2] Algérie Presse Service
https://www.aps.dz/sante-science-technologie. Consulté le 18 janvier 2024

[3] Institut National Du Cancer
https://www.e-cancer.fr/Patients-et-proches/Les-cancers/
Cancer-du-sein/Symptomes. Consulté le 20 janvier 2024

[4] American Cancer Society
https://www.cancer.org. Consulté le 21 janvier 2024

[5] Société Canadienne Du Cancer
https://cancer.ca/. Consulté le 23 janvier 2024

[6] Referentiels-Aristot
https://referentiels-aristot.com/.Consulté le 23 janvier 2024

[7] Imagerie Paris La Défense
https://radiologie-la-defense.fr/centre-paris/sein/.Consulté le 25 janvier 2024

[8] Aurélien Géron, Machine Learning avec Scikit-Learn, 2ème édition.

[9] IBM
https://www.ibm.com/fr-fr/topics/supervised-learning. Consulté le 2 février 2024

[10] DataScientest
https://datascientest.com/apprentissage-supervise. Consulté le 2 février 2024

[11] ALTERYX
https://www.alteryx.com/fr/glossary/supervised-vs-unsupervised-learning. Consulté le 3 février 2024

[12] AURÉLIEN GÉRON, Deep Learning avec Keras et TensorFlow, 2ème édition.

[13] FUTURA, Deluzarch, C. (14 octobre 2023). Deep Learning : qu'est-ce que c'est., https://www.futura-sciences.com/. Consulté le 8 février 2024

[14] FUTURA, Alizé Turpin. (09 novembre 2023). Algorithme Perceptron, Présentation et Fonctionnement,
https://www.jedha.co/formation-ia/algorithme-perceptron/.Consulté le 9 février 2024

[15] WIKIPÉDIA
https://fr.wikipedia.org/wiki/Fonctiond'activation. Consulté le 10 février 2024

[16] MATHWORKS
https://fr.mathworks.com/discovery/convolutional-neural-network.html. Consulté le 10 février 2024

[17] STANFORD.EDU
https://stanford.edu/~shervine/l/fr/teaching/cs-230/pense-bete-reseaux-neurones-convolutionnels. Consulté le 12 février 2024

[18] DATAANALYTICSPOST
https://dataanalyticspost.com/Lexique/svm. Consulté le 25 février 2024

[19] IBM
https://www.ibm.com/fr-fr/topics/random-forest. Consulté le 25 février 2024

[20] DATASCIENTEST
https://datascientest.com/knn. Consulté le 25 février 2024

[21] IBM
https://www.ibm.com/docs/fr/db2/11.5?topic=building-k-means-clustering. Consulté le 26 février 2024

[22] WIKIPÉDIA
https://fr.wikipedia.org/wiki/Algorithme_APriori. Consulté le 26 février 2024

[23] NAZARI, E., NADERI, H., TABADKANI, M. ET AL. BREAST CANCER PREDICTION USING DIFFERENT MACHINE LEARNING METHODS APPLYING

multi factors. J Cancer Res Clin Oncol 149, 17133–17146 (2023). https://doi.org/10.1007/s00432-023-05388-5

[24] David A. Omondiagbe et al 2019 IOP Conf. Ser. : Mater. Sci. Eng. 495 012033 DOI 10.1088/1757-899X/495/1/012033,

[25] Gardezi SJS, Elazab A, Lei B, Wang T Breast Cancer Detection and Diagnosis Using Mammographic Data : Systematic Review J Med Internet Res 2019 ;21(7) :e14464 doi : 10.2196/14464 PMID : 31350843 PMCID : 6688437

[26] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Torreão JRA, Broeders M, Gennaro G, Clauser P, Holen ÅS, Vollsæter M, Waade GG, Hofvind S, Houssami N. Performance of a Breast Cancer Detection AI Algorithm Using the Personal Performance in Mammographic Screening Scheme. Radiology. 2022 Oct ;305(1) :78-87. doi : 10.1148/radiol.220106. Epub 2022 Jun 28. PMID : 35764380.

[27] Abdulqader Mohammed, Mohammed Abdel Razek, Mohamed Eldosuky, Ahmed Sobhi Breast Cancer Detection Using Deep Learning Technique Based On Ultrasound Image 4 Dec 2023 doi : https://doi.org/10.48550/arXiv.2312.05261

[28] Jadoon, M. M., Zhang, Q., Haq, I. U., Butt, S., & Jadoon, A. (2017). Three-Class Mammogram Classification Based on Descriptive CNN Features. Volume 2017, Article ID 3640901. Queen Mary University of London, London, UK, & Faculty of Engineering and Technology, International Islamic University Islamabad, Islamabad, Pakistan. https://doi.org/10.1155/2017/3640901

[29] Antropova N, Huynh B, Giger Maryellen (2017) Performance compari son of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignant breast lesions on DCE-MRI. In : Proceedings of SPIE 10134, medical imag ing

[30] Teresa Araújo,Guilherme Aresta,Eduardo Castro,José Rouco,Paulo Aguiar,Catarina Eloy,António Polónia,Aurélio Campilho.https://doi.org/10.1371/journal.pone.0177544

[31] datascientest
https://acteurdemasante.lu/fr/cancer-du-sein/ Consulté le 18 mars 2024

[32] Wikipédia
https://fr.wikipedia.org/wiki/Mammographie Consulté le 19 mars 2024

[33] Institut National Du Cancer
https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Diagnostic/IRM Consulté le 22 mars 2024

[34] Doutriaux-Dumoulin, I. (2009, May 29). *Échographie mammaire et imagerie interventionnelle*. Séance organisée par la SOFMIS. Available online. https://doi.org/10.1016/S0221-0363(07)81105-X.

[35] EL SAN
https://www.elsan.care/fr/nos-equipements/pet-scan/ Consulté le 22 mars 2024

[36] analyticsvidhya, Kevin Kibe. (22 Jan, 2024). Deep Learning for Image Segmentation with TensorFlow,
http://surl.li/szlis. Consulté le 24 mars 2024

[37] David A. Omondiagbe et al 2019 IOP Conf. Ser. : Mater. Sci. Eng. 495 012033 DOI 10.1088/1757-899X/495/1/012033,

[38] Antropova N et al (2017) A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys 44(10) :5162–5171

[39] Antropova N, Huynh B, Giger M (2018) Recurrent neural networks for breast lesion classifcation based on DCE-MRIs. In : Proceedings of SPIE 10575, medical imaging 2018 : computer-aided diagno sis, 105752M. https://doi.org/10.1117/12.2293265

[40] praedictia
https://praedictia.com/page/apprentissage-profond/lhistoire-de-lapprentissage-profond.html

[41] wikipedia
https://fr.wikipedia.org/wiki/Anaconda_(distribution_Python).

[42] datascientest.
https://datascientest.com/numpy Consulté le 26 avril 2024

[43] Pandas
https://fr.wikipedia.org/wiki/Pandas Consulté le 28 avril 2024

[44] Matplotlib
https://fr.wikipedia.org/wiki/Matplotlib Consulté le 28 avril 2024

[45] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[46] Kaggle
https://www.kaggle.com/datasets/alaminbhuyan/breast-histopathology-images Consulté le 18 mai 2024

[47] datascientest
https://datascientest.com/quest-ce-que-le-modele-vgg.Consulté le 27 mai 2024

[48] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261-2269, doi : 10.1109/CVPR.2017.243.

[49] Mathworks
https://fr.mathworks.com/help/deeplearning/ref/resnet50.html Consulté le 22 mai 2024

[50] Lin, Ting-Yu ; Huang, Mei-Ling (2020), "Dataset of Breast mammography images with Masses", Mendeley Data, V2, doi : 10.17632/ywsbh3ndr8.2

[51] Taipy
https://docs.taipy.io/en/latest/ Consulté le 1 juin 2024

Résumé

Ce mémoire traite de l'application des techniques d'apprentissage automatique et de deep-learning pour la prédiction et le diagnostic du cancer du sein. L'objectif principal est de développer un modèle de machine learning capable de prédire le cancer du sein sur la base de données textuelles liées à des patients et de développer un modèle de deep-learning capable d'analyser des images histopathologiques afin de distinguer les tissus cancéreux des tissus sains. Nous avons utilisé, pour cela, le dataset Breast Cancer Winsconsin pour la partie prédiction et le dataset Breast Histopathology Images pour la partie classification/diagnostic. Nous avons expérimenté et évalué plusieurs techniques comme KNN, ANN, SVM, CNN etc... Les résultats que nous avons obtenus montrent un vrai potentiel de ces techniques pour la prédiction précoce et l'aide au diagnostic du cancer du sein. En complément, une application interactive a été développée en utilisant la bibliothèque Taipy, permettant aux cliniciens de télécharger des images et d'obtenir des prédictions en temps réel. Cette application facilite l'intégration des modèles d'intelligence artificielle dans la pratique clinique.

Mots Clés : Cancer du sein, Apprentissage automatique, Classification, Prédiction, Diagnostic médical, Réseaux de neurones, Deep learning.

Abstract

This report deals with the application of machine learning and deep-learning techniques for the prediction and diagnosis of breast cancer. The main objective is to develop a machine learning model capable of predicting breast cancer based on patient-related text data, and to develop a deep-learning model capable of analyzing histopathological images to distinguish cancerous from healthy tissue. We used the Breast Cancer Winsconsin dataset for prediction and the Breast Histopathology Images dataset for classification/diagnosis. We experimented with and evaluated several techniques such as KNN, ANN, SVM, CNN etc. The results we obtained show the real potential of these techniques for the early prediction and diagnosis of breast cancer. In addition, an interactive application has been developed using the Taipy library, enabling clinicians to download images and obtain predictions in real time. This application facilitates the integration of artificial intelligence models into clinical practice.