الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة أبي بكر بلقايد – تلمسان –
Université **Aboubakr Belkaïd**– **Tlemcen** –
Faculté de TECHNOLOGIE

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Génie Biomédical

Spécialité : Imagerie Médical

Par : **GUEDOUDOU Chourouk Hidayet**

Sujet

Comparaison des techniques d'explicabilité pour des méthodes de classification supervisées sur des images mammographiques

Soutenu publiquement, le **12 /06 /2025** , devant le jury composé de :

Mr GAOUAR Adil	MAA	Université de Tlemcen	Président
Mr TALEB Tariq	MCB	Université de Tlemcen	Examinateur
Mme KAZI TANI Lamia Fatiha	MCA	Université de Tlemcen	Encadrante

Année universitaire : **2024 /2025**

كـلـمـة شـكـر و تـقـديـر

الشكر أولا إلى الله عز و جل القائل في محكم كتابه العزيز

﴿ لَئِن شَكَرْتُمْ لَأَزِيدَنَّكُمْ ﴾

فشكرا لله سبحانه وتعالى على ما أسبغه علينا من نعم , و على تيسير السبيل , فله الحمد و الشكر في كل وقت وحين .
كما أتقدم بالشكر الجزيل للأستاذة **"كازي تاني لامية فتيحة"** لكل جهودها المبذولة في تسهيل و تيسير مهمة انجاز هذا البحث .

و اشكر الأستاذة **"حمزة الشريف سعاد "** على نصائحها المفيدة التي ساهمت انجاز بحثي بنجاح .
شكرا لمن جعل الله الجنة تحت إقدامها و التي احتضنني قلبها قبل يدها , التي سهلت لي الشدائد بدعائها
"والدتي"

شكرا **لعائلتي** على إيمانها و دعمها المتواصل .

شكرا لرفاق الخطوة الأولى و الخطوة ما قبل الأخيرة , إلى من كانوا سندا و عونا , وكنا نتقاسم التعب معا , نشارك الطموح و نتحدى الصعاب , إلى من ضحكنا معا و تعبنا معا , و دفعنا بعضنا البعض نحو الأفضل , ولولا وجودهم و روحهم الطيبة لما كانت هذه الرحلة ممتعة بهذا القدر
"أصدقائي".

و بكل احترام و تقدير , أتقدم بجزيل الشكر و الامتنان لأفراد اللجنة الكرام على وقتهم و جهودهم المبذولة في تقييم مشروعي و مناقشته .

إ هـد ا ء

ها أنا أقف اليوم على أعتاب حلمٍ طالما راودني، محمّلةً بفرح الإنجاز ودموع الامتنان، وما كان لهذا الطريق أن يُزهر لولا من كانوا النور فيه والسند على امتداده , فبهذه المناسبة اهدي تخرجي

إلى أمي الغالية،

إلى من أثريت روحي بالإيمان، وكنتِ لي ملجأً آمنًا في كل لحظة ضعف، دعاؤكِ كان نوري في العتمة، وصبركِ كان دافعًا لأكمل. كنتِ لي الأم والأب معًا، تحملتِ الكثير بدوني أن أشعر بنقص أو ضعف، فأعطيتني من الحنان والقوة ما يكفيني طوال حياتي. هذا النجاح لكِ أولًا، فأنتِ من علّمتني أن لا مستحيل مع العزيمة والدعاء.

إلى أختي الحبيبة،

إلى من كانت ابتسامتكِ بلسمًا، وقلبكِ احتواني بلا شروط. كنتِ شريكتي في الدرب، ورفيقة روحي... وجودكِ بجانبي جعل كل شيء أهون.

إلى عائلتي الكريمة،

لكم مني كل الحب، فأنتم جذوري، ومنكم استمددت قوتي، وفي دعواتكم كانت راحتي. أنتم البركة التي لا تزول، والنبض الذي يمدّني بالحياة.

إلى صديقاتي العزيزات،خلود ومنال،

كنتنّ الضوء في ليالي التعب، واليد التي ساندتني حين تعثّرت. ما أجمل الحياة بقلوبٍ مثلكم، دافئة، صادقة، لا تعرف إلا الوفاء،أنتما أكثر من مجرد أسماء في حياتي... أنتما ذاكرة مليئة بالمحبة، وصوت داعم في كل منعطف. بوجودكما، كان للحياة طعمٌ أجمل.

وإلى مشرفتي الفاضلة،

كل الشكر والتقدير لكِ، فقد كنتِ المعينة في دربي، وكنتِ نورًا وجّهني حين تاهت الخطى. لكِ بصمة خالدة في هذا الإنجاز.

إلى كل من آمن بي، ورفع يده بدعوة، أو قدّم لي كلمة طيبة... هذا التخرّج ليس إنجازي وحدي، بل هو ثمرة قلوبكم التي أحبّتني بصدق .

Résumé

Dans le contexte où les outils automatisés prennent une place de plus en plus importante dans le secteur médical, en particulier pour l'analyse d'images. Cependant, il reste essentiel que les décisions prises par ces systèmes soient compréhensibles et transparentes pour les professionnels de santé.

Ce mémoire se concentre sur la comparaison de différentes méthodes permettant d'expliquer et d'interpréter les résultats obtenus par des modèles de classification supervisée appliqués à des images de mammographies. L'objectif est de faciliter la compréhension des prédictions afin de mieux accompagner les médecins dans leur diagnostic.

Pour ce faire, des réseaux de neurones convolutifs ont été utilisés pour identifier la présence de cancer du sein à partir de la base de données ***''Final-Final-RSNA-Breast-Cancer-Dataset ''***. Plusieurs techniques d'explication, comme LIME, SHAP et Grad-CAM, ont été mises en œuvre pour mettre en lumière les parties de l'image ayant le plus influencé la décision du modèle, rendant ainsi les résultats plus accessibles.

Les tests réalisés ont montré des performances solides, avec une précision de classification(accuracy) de 99,51 %, une perte faible lors de l'apprentissage, et une précision spécifique de 99,63%. Ces résultats, associés à des visualisations explicatives claires, démontrent l'intérêt d'intégrer ces méthodes d'interprétation dans les systèmes d'aide au diagnostic. Ce travail souligne l'importance de la transparence pour renforcer la confiance des utilisateurs et garantir une utilisation responsable et éthique des technologies dans le domaine de la santé.

Mots-clefs : Intelligence artificielle(IA) ,Apprentissage profond ,Réseaux de neurones convolutifs (CNN),XAI (eXplainable Artificial Intelligence),Cancer du sein,Mammographie

ملخص

في سياقٍ تتزايد فيه أهمية الأدوات الآلية في القطاع الطبي، لا سيما في تحليل الصور، يبقى من الضروري أن تكون القرارات التي تتخذها هذه الأنظمة مفهومة وشفافة لمقدمي الرعاية الصحية.تركز هذه الأطروحة على مقارنة مختلف أساليب شرح وتفسير نتائج نماذج التصنيف المُشرف عليها والمُطبقة على صور تصوير الثدي بالأشعة السينية. الهدف هو تسهيل فهم التنبؤات لدعم الأطباء بشكل أفضل في تشخيصهم.

ولهذا الغرض، استُخدمت الشبكات العصبية التلافيفية لتحديد وجود سرطان الثدي من قاعدة بيانات-***''Final-Final***
RSNA-Breast-Cancer-Dataset ''.وطُبقت العديد من تقنيات الشرح، مثل LIME و SHAPوGrad-CAM، لتسليط الضوء على أجزاء الصورة الأكثر تأثيرًا على قرار النموذج، مما يجعل النتائج أكثر سهولة في الوصول إليها. أظهرت الاختبارات التي أُجريت أداءً قويًا، بدقة تصنيف بلغت 99.51%، وخسارة منخفضة أثناء التدريب، ودقة نوعية بلغت 99.63%. تُظهر هذه النتائج، إلى جانب التصورات التوضيحية الواضحة، أهمية دمج أساليب التفسير هذه في أنظمة دعم التشخيص. ويُسلّط هذا العمل الضوء على أهمية الشفافية لتعزيز ثقة المستخدم وضمان الاستخدام المسؤول والأخلاقي للتقنيات في مجال الرعاية الصحية.

الكلمات المفتاحية: الذكاء الاصطناعي (AI)، التعلم العميق، الشبكات العصبية التلافيفية ((CNN، XAI (الذكاء الاصطناعي القابل للتفسير)، سرطان الثدي، التصوير الشعاعي للثدي.

Abstract

In a context where automated tools are playing an increasingly important role in the medical sector, particularly for image analysis, it remains essential that the decisions made by these systems be understandable and transparent for healthcare professionals.

This thesis focuses on comparing different methods for explaining and interpreting the results obtained by supervised classification models applied to mammogram images. The goal is to facilitate understanding of the predictions in order to better support physicians in their diagnosis.

To this end, convolutional neural networks were used to identify the presence of breast cancer from the **''Final-Final-RSNA-Breast-Cancer-Dataset''** database. Several explanation techniques, such as LIME, SHAP, and Grad-CAM, were implemented to highlight the parts of the image that most influenced the model's decision, thus making the results more accessible.

The tests carried out showed solid performance, with a classification accuracy of 99.51%, a low loss during training, and a specific accuracy of 99.63%. These results, combined with clear explanatory visualizations, demonstrate the value of integrating these interpretation methods into diagnostic support systems. This work highlights the importance of transparency to strengthen user trust and ensure responsible and ethical use of technologies in the healthcare field.

Keywords: Artificial Intelligence (AI), Deep Learning, Convolutional Neural Networks (CNN), XAI (eXplainable Artificial Intelligence), Breast Cancer, Mammography.

Table de matières

Liste des figures

Liste des tableaux

Abréviations et acronymes

IA	**I**ntelligence **A**rtificielle .
ML	**M**achine **L**earning .
XAI	e**X**plainable **A**rtificial **I**ntelligence .
SVM	**S**upport **V**ector **M**achine .
SSL	**S**emi **S**upervise **L**earning .
DL	**D**eep **L**earning .
CNN	**C**onvolutional **N**eural **M**achine .
VGG	**V**isual **G**eometry **G**roup .
ResNet	**R**esidual **N**etwork .
RL	**R**egression **L**ogistique .
KNN	**K**-**N**earest **N**eighbors .
GAM	**G**eneralized **A**dditive **M**odel .
RF	**R**andom **F**orest .
MNN	**M**ulti-**L**ayer **N**eural **N**etwork .
RNN	**R**ecurrent **N**eural **N**etwork .
LIME	**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations .
G-REX	**G**enitic **R**ule **EX**traction .
CNF	**C**onjuctive **N**ormal Form .
DNF	**D**isjunctive **N**ormal **F**orm .
SHAP	**SH**apley **A**dditive ex**P**lanations .
QII	**Q**uantitative **I**nput **I**nfluence .
ASTRID	**A**utomatic **STR**ucture **ID**entification method .
ICE	**I**ndividual **C**onditional **E**xpectation .
PD	**P**artial **D**ependence .
RE	**R**écepteur aux **Œ**strogènes .
RP	**R**écepteur aux **P**rogestérone .
CO	**C**ontraceptifs **O**raux .
THS	**T**raitement **H**ormonal **S**ubstitutifs .
IRM	**I**magerie par **R**ésonance **M**agnétique .

Introduction générale

Aujourd'hui, l'intelligence artificielle (IA) joue un rôle croissant dans le domaine médical, en particulier en imagerie, où elle permet d'automatiser et de fiabiliser de nombreuses tâches complexes. Grâce à l'apprentissage automatique, et plus précisément aux modèles de classification supervisée, il est désormais possible d'analyser des images médicales avec une précision impressionnante. Dans le cas du dépistage du cancer du sein, les mammographies représentent une source précieuse d'informations, et les modèles d'IA sont capables de détecter des signes précoces de la maladie, parfois difficiles à percevoir pour l'œil humain.
Cependant, cette avancée technologique s'accompagne d'un défi de taille : comprendre le fonctionnement de ces modèles. Bien qu'efficaces, ils agissent souvent comme des "boîtes noires", produisant des résultats sans que l'on sache réellement ce qui a motivé la décision. Or, en médecine, il ne suffit pas de fournir une réponse automatisée : il faut pouvoir expliquer cette réponse, la justifier, et la confronter à l'avis clinique. Sans cette transparence, il est difficile pour les médecins de faire confiance à ces outils, et donc de les intégrer réellement dans leur pratique.

C'est dans cette optique que se développent les méthodes d'explicabilité, connues sous le nom de XAI (eXplainable Artificial Intelligence). Leur but est de rendre les prédictions des modèles d'IA plus compréhensibles, notamment en visualisant les éléments d'une image qui ont influencé la décision. Elles représentent une passerelle entre la performance algorithmique et la validation humaine, en aidant les professionnels de santé à interpréter les résultats générés par la machine.

Cela soulève alors une question essentielle, au cœur de ce projet de fin d'études : quelles méthodes d'explicabilité sont les plus pertinentes pour accompagner les modèles de classification supervisée appliqués à des images de mammographies, tout en restant compréhensibles, fiables et utiles aux médecins ?

Pour répondre à cette problématique, plusieurs étapes ont été nécessaires. Le travail a d'abord commencé par une étude théorique approfondie de l'intelligence artificielle appliquée à l'imagerie médicale, avec un focus sur les enjeux de l'interprétabilité. Ensuite, des modèles de

classification ont été conçus et testés sur des mammographies, avant d'y intégrer différentes méthodes explicatives comme LIME, SHAP ou Grad-CAM. Enfin, une phase d'analyse comparative a permis d'évaluer la qualité des explications produites, en mesurant leur pertinence et leur potentiel à être utilisés dans un contexte clinique.

À travers cette démarche, l'objectif est de proposer une contribution concrète à la médecine de demain : des outils d'aide à la décision qui ne soient pas seulement performants, mais aussi transparents et acceptés par les professionnels de santé.

Chapitre

I

L'intelligence artificielle et l'IA explicable

1. Introduction

L'intelligence artificielle (IA), et plus spécifiquement le machine learning supervisé, joue aujourd'hui un rôle déterminant dans de nombreux domaines, dont celui de la médecine, en permettant des avancées significatives dans l'automatisation des diagnostics et l'amélioration de la précision des décisions cliniques.Dans le contexte de l'analyse d'images médicales, comme les mammographies, a bénéficié de ces avancées en permettant d'automatiser la détection des signes précoces du cancer du sein avec une précision remarquable Cependant, bien que ces modèles offrent des résultats impressionnants en termes de précision, une limite persiste : leur manque d'explicabilité. En effet, bien que les prédictions soient fiables, comprendre les raisons qui sous-tendent une décision reste souvent opaque, ce qui soulève des préoccupations quant à la transparence et à la confiance envers ces systèmes, en particulier dans des applications critiques comme la santé.

C'est ici qu'intervient l'IA explicable, ou XAI (eXplainable Artificial Intelligence), un domaine qui vise à rendre les modèles d'IA plus compréhensibles et transparents. Dans un contexte médical, cette explicabilité est essentiel : elle permet aux professionnels de santé de mieux interpréter les résultats, renforcer la confiance et garantir des décisions cliniques responsables et éthiques.

Ce chapitre explore les différentes approches d'explicabilité appliquées aux modèles de machine learning supervisé, avec un focus particulier sur l'analyse d'images médicales. Nous verrons comment ces méthodes contribuent à rendre les prédictions plus transparentes et compréhensibles, et en quoi elles sont essentielles pour le diagnostic du cancer du sein.

2. L'intelligence artificielle

L'intelligence artificielle (IA) désigne un ensemble d'algorithmes capables de traiter des données afin de simuler des tâches cognitives humaines telles que la prise de décision ou la résolution de problèmes. Un système intelligent intègre à la fois une composante matérielle (dispositifs électroniques) et une composante logicielle (programmes de traitement d'informations). L'IA s'inscrit à l'intersection de plusieurs disciplines comme l'informatique, les mathématiques et les sciences cognitives, et repose sur des approches variées telles que l'apprentissage automatique, les réseaux de neurones ou la représentation symbolique. (Zara.I, 2019)

2.1. Apprentissage Automatique

L'apprentissage automatique (ou apprentissage machine, machine learning) est un sous domaine de l'IA qui s'intéresse en particulier aux capacités d'apprentissage. Le principe est de reproduire un comportement non pas en le programmant "à la main" dans un ordinateur, mais en concevant un système plus général capable d'apprendre à partir d'exemples à résoudre votre problème. (Barka.Z, 2022)

2.1.1. Les types de l'apprentissage automatique

2.1.1.1. Apprentissage supervisé

La forme la plus courante d'apprentissage automatique est l'apprentissage supervisé. L'apprentissage supervisé est une méthode de transformation d'un ensemble de données en un autre, le programme est entraîné sur un ensemble prédéfini d'exemples d'entraînement, ce qui facilite ensuite sa capacité à parvenir à une conclusion précise lorsque de nouvelles données sont fournies. Les algorithmes de classification supervisée de ML sont : Forest Random, Decision Trees, Logistic Regression, et le plus connu est SVM Support Vector Machines. (Barka.Z, 2023)

2.1.1.2. Apprentissage non supervisé

L'apprentissage non supervisé, également connu sous le nom d'apprentissage à partir d'observations, partage une propriété commune avec l'apprentissage supervisé : il transforme un ensemble de données en un autre. Mais l'ensemble de données dans lequel il se transforme n'est pas connu ou compris auparavant. Contrairement à l'apprentissage supervisé, il ne se nourrira quant à lui que d'exemples, et créera lui-même les classes qu'il jugera les plus judicieuses (clustering) ou des règles d'association (algorithmes Apriori). L'algorithme K-mean (Kmeans) permet de comprendre facilement le concept de classification non supervisée (Barka.Z, 2023) .

2.1.1.3. Apprentissage semi supervisé

Le protocole SSL (Semi Supervise Learning) est spécialement conçu pour les secteurs d'application où les données non étiquetées sont fréquentes, comme le traitement des images, la collecte d'informations et la bioinformatique . La technologie SSL se situe à mi-chemin entre l'apprentissage supervisé et non supervisé, ce qui signifie que l'ensemble des données est séparé en étiquettes et sans étiquettes. L'apprentissage semi supervisé consiste à utiliser des données non-annotées afin de compléter l'apprentissage supervisé (Lazhar.M, 2022) .

2.2. Apprentissage profond

L'apprentissage profonde (Deep learning ou DL) appartient à une classe de techniques d'apprentissage automatique (machine learning ou ML) *(Figure 1),* il obtient un grand succès dans de nombreuses tâches de l'intelligence artificielle (IA) par rapport aux algorithmes de ML classiques. Les architectures des modèles profondes sont relativement récentes où de nombreuses étapes de traitement non linéaire de l'information sont exploitées, dans lesquelles les informations sont traitées en couches hiérarchiques, chacune recevant et interprétant les informations de la couche précédente pour l'apprentissage des représentations de données (Barka.Z, 2022) .

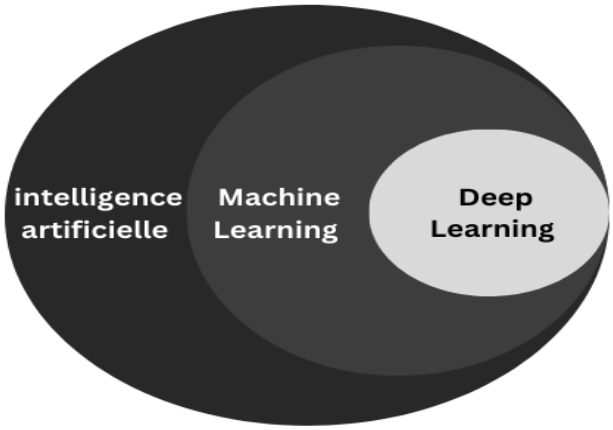

Figure 1.1. La relation entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond. (Barka.Z, 2022)

Généralement, l'architecture des réseaux profondes est organisée en couches de neurones pour n'importe quel type de ces réseaux; une Couche d'entrée (Input Layer), une ou plusieurs Couches cachées (Hidden Layers) et une Couche de sortie (Output Layer) (Barka.Z, 2022).
Chaque paire de couches voisines est connectée. Les connexions entre eux appelées poids (Weights). Les "neurones" d'une même couche généralement appelés "nœuds" n'ont aucune association, la figure 1.2illustré une architecture standard d'un modèle de réseau de neurones profond (Barka.Z, 2022).
L'apprentissage profond se présente comme un système de calcul avancé, il est constitué d'une variété de techniques issues du domaine de l'apprentissage automatique qui utilisent un déluge de neurones (nœuds) non linéaires disposés en plusieurs couches de traitement qui extraient et convertissent des valeurs de variables d'entité à partir du vecteur d'entrée pour créer plusieurs niveaux d'abstraction afin de représenter les données (Barka.Z, 2022).

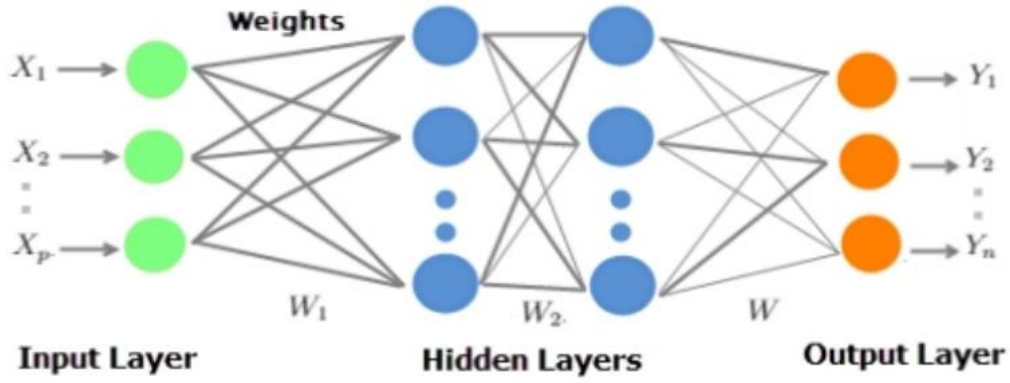

Figure 1.2.L'architecture d'un modèle Deep Learning. (Barka.Z, 2022)

Parmi les modèles les plus performants et les plus couramment utilisés, notamment pour des tâches telles que la classification d'images médicales.

2.2.1. Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutionnels proposés initialement par Le Cun. Ce choix a été motivé principalement par ce qu'il intègre implicitement une phase d'extraction de caractéristiques et il a été utilisé avec succès dans de nombreuses applications. Ils sont réputés pour leur robustesse aux faibles variations d'entrée, le faible taux de prétraitement nécessaires à leur fonctionnement **(Nihad.B et al, 2019)**.

Le CNN est un réseau de neurone multicouche qui est spécialisés dans des tâches de reconnaissance de forme. Ces réseaux ont été inspirés par les travaux de Hubel et Wiesel sur le cortex visuel chez les mammifères qui combine trois idées principales :

-les champs récepteurs locaux.

-les poids partagés.

-le sous-échantillonnage.

L'architecture de CNN repose sur plusieurs réseaux de neurones profonds consistant en une succession de couches de convolution et d'agrégation (pooling) est dédié à l'extraction automatique de caractéristiques, tandis que la seconde partie, composée de couches de neurones complètement connectés, est dédiée à la classification **(Nihad.B et al, 2019)**.

Chaque cellule des couches de convolution est connectée à un ensemble de cellules regroupées dans un voisinage rectangulaire sur la couche précédente. Les champs récepteurs locaux permettent d'extraire des caractéristiques basiques. Les couches sont dites « à convolution » car les poids sont partagés et chaque cellule de la couche réalise la même combinaison linéaire (avant d'appliquer la fonction sigmoïde) qui peut être vue comme une simple convolution. Ces caractéristiques sont alors combinées à la couche suivante afin dedétecter des caractéristiques de plus haut niveau **(Nihad.B et al, 2019)**.

Entre deux phases d'extraction de caractéristiques, le réseau réduit la résolution de la carte des caractéristiques par un moyen de sous-échantillonnage. Cette réduction se justifie à deux titres : diminuer la taille de la couche et apporter de la robustesse par rapport aux faibles distorsions **(Nihad.B et al, 2019)**.

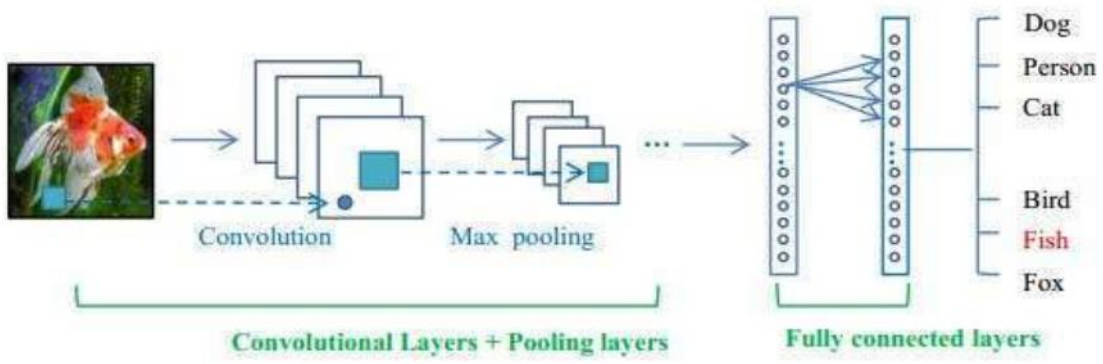

Figure 1.3 . L'architecture d'un réseau de neurone convolutif. (Nihad.B et al, 2019)**.**

2.2.1.1. Couche de convolution

La convolution est une opération mathématique comme l'addition et la multiplication, il est très utile de simplifier des équations plus complexes, cette opération est largement utilisée dans le traitement du signal numérique **(Nihad.B et al, 2019)**.

Lorsque on applique la convolution aux le traitement d'image, on convoler (combiner) l'image d'entrée avec une sous-région de cette image (filtre). Le filtre est aussi connu sous le nom du noyau de convolution, il consiste en des poids de cette sous-région. La sortie de cette couche est l'image entrée avec des modifications qui est souvent appelée une carte de caractéristique (feature Map) **(Nihad.B et al, 2019)**.

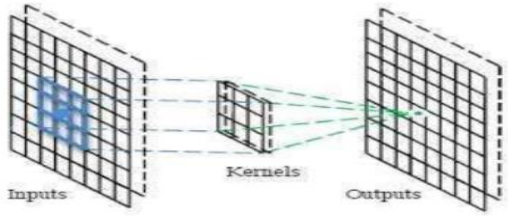

Figure 1.4.L'opération de convolution. (Nihad.B et al, 2019)**.**

En terme mathématique, Une couche de convolution Ci (couche i du réseau) est paramétrée par son nombre N de cartes de convolution$\boldsymbol{M_j^i}$ (j ∈ {1,..., N}), la taille des noyaux de convolution K x × Ky(souvent carrée), et le schéma de connexion à la couche précédente$\boldsymbol{L^i}$ **(Nihad.B et al, 2019)**.

Chaque carte de convolution $\boldsymbol{M_j^i}$est le résultat d'une somme de convolution des cartes de la couche précédente$\boldsymbol{M_j^i}$ par son noyau de convolution respectif. Un biais $\boldsymbol{b_j^i}$est ensuite ajouté et le résultat est passe à une fonction de transfert non-linéaire $\boldsymbol{\varphi}$. Dans le cas d'une carte complètement connectée aux cartes de la couche précédente, le résultat est alors calculé par **(Nihad.B et al, 2019)**:

$$\boldsymbol{M_j^i = \varphi\,(b_j^i + \sum_{n=1}^{n} M_j^{i-1} K_j^i\,)}$$

2.2.1.2. Couche de sous-échantillonnage (Pooling)

Dans les architectures classiques de réseaux de neurones convolutionnels, les couches de convolution sont suivies par des couches de sous échantillonnage (couche d'agrégation). Cette dernière réduit la taille des cartes de caractéristique pour but de diminuer la taille de paramètre, et renvoie les valeurs maximales des régions rectangulaires de son entrée **(Nihad.B et al, 2019)**.

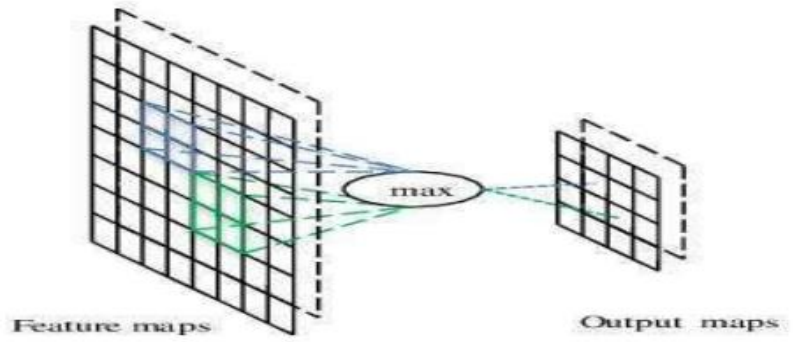

Figure 1.5 .L'opération de sous échantillonnage. (Nihad.B et al, 2019)

2.2.1.3. Couche entièrement connectée

Les paramètres des couches de convolution et de max agrégation sont choisis de sorte que les cartes d'activation de la dernière couche soient de taille 1, ce qui résulte en un vecteur 1D d'attributs. Des couches classiques complètement connectées composées de neurones sont alors ajoutées au réseau pour réaliser la classification **(Nihad.B et al, 2019)**.

La dernière couche, dans le cas d'un apprentissage supervisé, contient autant de neurones que de classes désirées. Cette dernière couche contient N neurones(nombre des classes dans la base), et une fonction d'activation de type sigmoïde est utilisée afin d'obtenir des probabilités d'appartenance à chaque classe **(Nihad.B et al, 2019)**.

2.2.2. Les architectures des CNN

Ces dernières années, les modèles d'apprentissage profond ont révolutionné la classification des images, offrant une robustesse supérieure aux méthodes traditionnelles. Cette section mettra en lumière les modèles les plus reconnus dans la littérature, en particulier VGG ,ResNet et EfficentNet .

2.2.2.1. Res-Net

C'est l'un des réseaux de neurones profonds les plus puissants qui a obtenu d'excellents résultats de performances dans le défi de classification ILSVRC 2015. Res-Net a réalisé d'excellentes performances de généralisation sur d'autres tâches de reconnaissance et a Remporté la première place sur la détection Image-Net, la localisation Image-Net, la détection COCO et la segmentation COCO dans les concours ILSVRC et COCO 2015. Il existe de

nombreuses variables de l'architecture Res-Net, c'est-à-dire le même concept mais avec un nombre de couches différent. Nous avons **Res-Net-18, Res-Net-34, Res-Net-50, ResNet-101, Res-Net-110, Res-Net-152, Res-Net-164, Res-Net-1202 (Sari.Y, 2023)** .

2.2.2.2. VGG-16

VGG-16 est un modèle qui a atteint une précision de 92,7% dans le top 5 des tests dans Image-Net, qui est un ensemble des données de plus de 14 millions d'images appartenant à 1000 classes. C'était l'un des célèbres modèles soumis à la conférence ILSVRC- 2014 .

VGG-16 est constitué de 16 couches (13 couches enveloppantes et 3 couches entièrement connectées) **(Sari.Y, 2023)**.

2.2.2.3. Efficentnet

EfficientNet est une famille de réseaux de neurones convolutifs (CNN) qui vise à atteindre des performances élevées avec moins de ressources de calcul par rapport aux architectures précédentes. Elle a été présentée par Mingxing Tan et Quoc V **(Sharma.S, 2023)**. Le de Google Research dans leur article de 2019 « EfficientNet : Repenser la mise à l'échelle des modèles pour les réseaux de neurones convolutifs ». L'idée principale d'EfficientNet est une nouvelle méthode de mise à l'échelle qui adapte uniformément toutes les dimensions de profondeur, de largeur et de résolution à l'aide d'un coefficient composé **(Sharma.S, 2023)**.

Elle démarre avec une couche d'entrée (*stem*),composée d'une convolution classique, d'une normalisation et d'une activation ReLU6 . Son cœur est constitué de plusieurs blocs appelés **MBConv**. Ces blocs combinent des convolutions séparables en profondeur (*depthwise separable convolutions*) avec des couches squeeze-and-excitation. Ce mécanisme permet au réseau de mieux capturer les caractéristiques importantes tout en réduisant le nombre total de paramètres, ce qui le rend plusrapide et plus léger , Le réseau se termine avec la tête(*head*) qui compose d'une dernière couche de convolution, suivie d'un global average pooling, puis d'une couche entièrement connectéepour la classification finale **(Sharma.S, 2023)**.

3. Explicabilité de l'IA

3.1. Définition

L'IA explicable est un ensemble de processus et de méthodes qui permettent aux utilisateurs de comprendre et de se fier aux résultats et aux informations générés par les algorithmes de machine learning (ML) de l'IA. Les explications qui accompagnent les résultats de l'IA/ML peuvent être destinées aux utilisateurs, aux opérateurs ou aux développeurs. Elles sont conçues pour résoudre des problèmes et relever des défis tels que l'adoption par les utilisateurs, la gouvernance et le développement de système (Juniper.Networks, sd.) .

3.2.L'objectif de l'IA explicable

Les objectifs principaux de l'IA explicables :

- **Fiabilité :** C'est la capacité d'un modèle à donner des résultats cohérents et prévisibles. Cela dit, même si un modèle est fiable, ça ne veut pas forcément dire qu'on comprend comment il fonctionne. Et en pratique, évaluer cette fiabilité n'est pas toujours simple.

- **Causalité** : Un des grands enjeux de l'explicabilité, c'est de mieux comprendre les liens de cause à effet entre les variables. Les modèles classiques repèrent surtout des corrélations, mais l'explicabilité peut aider à détecter ou confirmer des relations causales.

- **Transférabilité :**Lorsqu'on comprend bien un modèle, il est plus facile de l'adapter à d'autres situations. L'explicabilité permet justement de voir dans quels cas un modèle peut (ou non) être réutilisé, en évitant les mauvaises surprises.

- **Informative :**Un modèle explicable doit donner assez d'informations pour que ses prédictions soient compréhensibles. Sinon, on risque de mal les interpréter, voire de prendre de mauvaises décisions.

- **Confiance :**Quand on comprend comment un modèle prend ses décisions, on a naturellement plus confiance en lui. C'est particulièrement important dans des domaines comme la santé, où les décisions peuvent avoir de lourdes conséquences.

- **Équité** L'explicabilité permet de repérer des traitements injustes ou des biais dans les données. C'est une façon de s'assurer que les décisions prises par l'IA sont éthiques et équitables.

- **Accessibilité :**Plus un modèle est explicable, plus il est facile à utiliser pour des personnes qui ne sont pas expertes en intelligence artificielle. Cela rend les technologies plus inclusives et favorise leur adoption.

- **Interactivité :**Certains systèmes permettent aux utilisateurs d'interagir avec le modèle, par exemple pour ajuster des paramètres ou poser des questions. Cela aide à mieux comprendre le fonctionnement du modèle et à en tirer le meilleur.

- **Conscience de la vie privée** : Expliquer un modèle peut, parfois, révéler des informations sensibles. Il faut donc trouver un bon équilibre entre transparence et respect de la confidentialité des données.

3.3.Catégories de méthodes

Les méthodes peuvent être regroupées par portée en celles fournissant des **explications globales** de l'ensemble du système et celles fournissant des **explications locales** d'une seule prédiction. Les techniques peuvent également être regroupées selon qu'elles sont **agnostiques au modèle** ou **spécifiques au modèle**. De plus, les techniques peuvent être divisées en méthodes d'explicabilité **ante-hoc** et **post-hoc** (Mooney, M, A, D, W, & G, 2011).

3.3.1. Méthodes locales ou globales

Les **explications globales** facilitent la compréhension du comportement et du raisonnement de l'ensemble du modèle conduisant à des résultats attendus. Pour **lesexplications locales**, les raisons d'une seule prédiction sont fournies pour justifier pourquoi le modèle a pris une décision spécifique pour cette instance (Mooney, M, A, D, W, & G, 2011).

3.3.2. Méthodes agnostiques ou spécifiques au modèle

Agnostiques au modèle c'est-à-dire qu'elles peuvent être appliquées à n'importe quel algorithme d'apprentissage automatique ou **spécifiques au modèle** c'est-à-dire qu'elles ne peuvent être appliquées qu'à un algorithme d'apprentissage automatique spécifique (Mooney, M, A, D, W, & G, 2011).

3.3.3. Méthodes ante-hoc ou post-hoc

Les méthodes d'explication ante-hoc, intègrent l'interprétabilité directement dans le modèle d'apprentissage. Cette interprétabilité intégrée peut être soit partielle au niveau local, par exemple, en utilisant des mécanismes d'attention pour indiquer quels attributs d'entrée sont pertinents pour les prédictions du modèle, soit globale. Un exemple bien connu de modèle interprétable est le modèle de régression linéaire : les coefficients associés à chaque attribut indiquent son influence. Cependant, un inconvénient potentiel de l'interprétabilité intégrée est que son incorporation dans le modèle peut limiter son architecture et potentiellement influencer ses performances (Bonia.L, 2023).

Les méthodes d'explication post-hoc fournissent des explications pour un modèle d'apprentissage automatique déjà entraîné. Elles peuvent donner une explication globale du raisonnement du modèle ou générer des explications locales pour expliquer une instance spécifique. Par exemple, les cartes de chaleur peuvent mettre en évidence les parties importantes d'une image pour la prédiction du modèle, les scores d'importance peuvent indiquer les mots pertinents pour la classification d'un texte, et des règles de décision sous forme de clauses ou de termes logiques peuvent également être utilisées pour expliquer. Une explication globale post-hoc pourrait prendre la forme d'un ensemble de règles qui reproduit le raisonnement du modèle et agit comme un substitut pour la boîte noire (Bonia.L, 2023).

4. Les méthodes XAI existantes :

Les algorithmes d'apprentissage automatique présentent des niveaux de performance variables et sont généralement moins précis que les approches basées sur l'apprentissage profond (Deep Learning). Ainsi, le choix de l'algorithme à utiliser dépend souvent d'un compromis entre la performance du modèle et sa capacité à fournir des explications compréhensibles (explicabilité). Comme illustré dans la Figure 1.6, différentes approches d'explicabilité ont été développées pour rendre plus transparentes les décisions prises par les modèles d'intelligence artificielle.

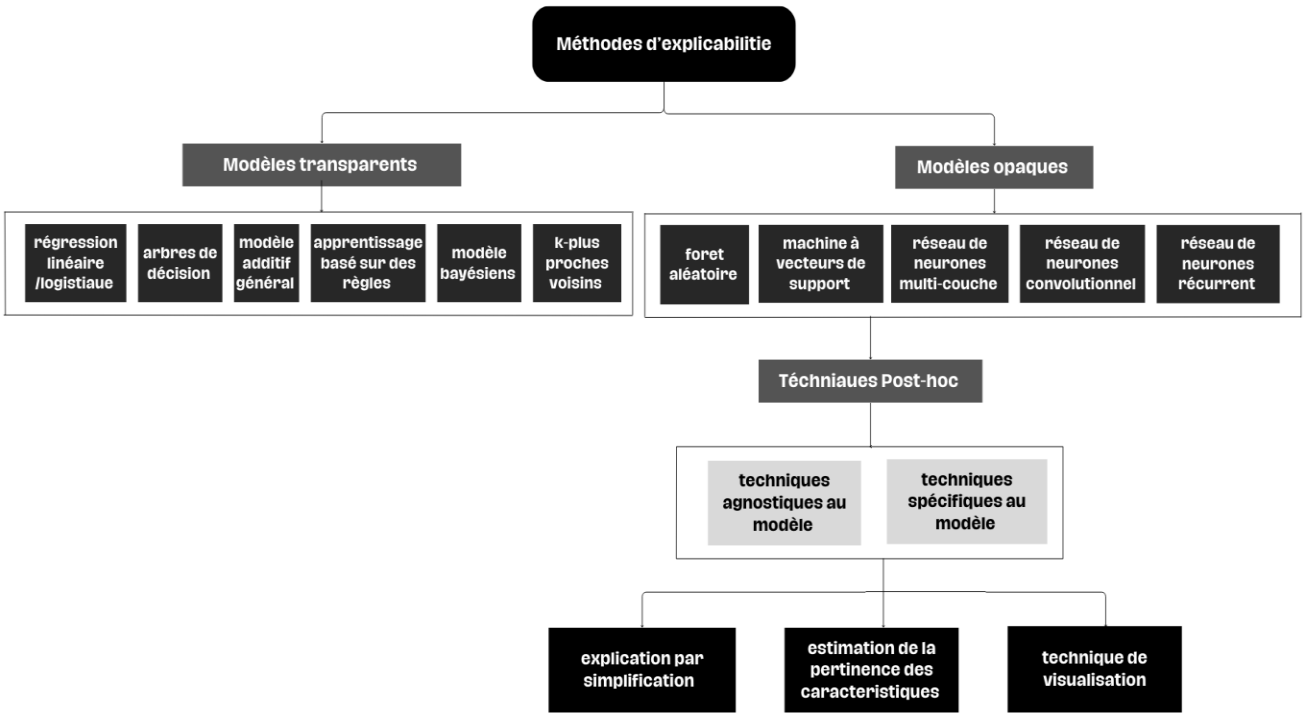

Figure 1.6. Un Schéma Hiérarchique : Comment Expliquer l'Intelligence Artificielle.

Cette figure met en lumière les différentes stratégies développées pour rendre les modèles d'intelligence artificielle plus interprétables. Les modèles simples, dits intrinsèquement explicables, offrent une transparence directe grâce à leur structure facilement compréhensible. En revanche, les modèles plus complexes, comme ceux issus du deep learning, nécessitent des approches dites post-hoc. Ces méthodes interviennent après l'entraînement du modèle afin de mieux comprendre ses décisions. Elles peuvent consister à analyser l'influence des variables d'entrée, à générer des visualisations des zones pertinentes, ou encore à approximer le comportement du modèle par une version simplifiée. Ces techniques permettent ainsi de lever partiellement le voile sur le fonctionnement interne des modèles, tout en préservant leur performance.

4.1.Les modèles transparents en ML

Est un modèle qui est expliqué pendant son entraînement. C'est-à-dire que l'architecture intrinsèque des modèles satisfait l'une des trois dimensions de transparence.il incluent :

4.1.1. Régression linéaire/logistique

La régression logistique (RL) est un modèle de classification utilisé pour prédire une variable dépendante (catégorie) qui est dichotomique (binaire). Cependant, lorsque la variable dépendante est continue, la régression linéaire en serait l'homonyme. Ce modèle repose sur l'hypothèse de dépendance linéaire entre les prédicteurs et les variables prédites, ce qui empêche une adaptation flexible aux données. Cette raison spécifique (la rigidité du modèle) est celle qui maintient le modèle sous l'égide des méthodes transparentes. Cependant, l'explicabilité est liée à un certain public, ce qui fait qu'un modèle peut tomber sous les deux catégories selon qui est censé l'interpréter. De cette manière, la régression logistique et la régression linéaire, bien qu'elles répondent clairement aux caractéristiques des modèles transparents (transparence algorithmique, décomposabilité et simulabilité), peuvent également nécessiter des techniques d'explicabilité post-hoc (principalement, la visualisation), en particulier lorsque le modèle doit être expliqué à des publics non experts (Alejandro.B et al, 2019).

4.1.2. Arbres de décision

Les arbres de décision sont un autre exemple de modèle qui peut facilement satisfaire toutes les contraintes de transparence. Les arbres de décision sont des structures hiérarchiques de prise de décision utilisées pour résoudre des problèmes de régression et de classification. Dans leur forme la plus simple, les arbres de décision sont des modèles simulables. Cependant, leurs propriétés peuvent les rendre décomposables ou algorithmiquement transparents. Les arbres de décision ont toujours oscillé entre les différentes catégories de modèles transparents. Leur utilisation a été étroitement liée aux contextes de prise de décision, ce qui explique pourquoi leur complexité et leur compréhension ont toujours été considérées comme des enjeux primordiaux. Une preuve de cette pertinence peut être trouvée dans l'essor des contributions à la littérature traitant de la simplification et de la génération des arbres de décision. Comme mentionné précédemment, bien qu'ils soient capables de s'adapter à chaque catégorie au sein des modèles transparents, les caractéristiques individuelles des arbres de décision peuvent les amener à appartenir à la catégorie des modèles algorithmiquement transparents. Un arbre de décision simulable est celui qui est gérable par un utilisateur humain. Cela signifie que sa taille est relativement petite et que le nombre de caractéristiques et leur signification sont facilement compréhensibles. Une augmentation de la taille transforme le modèle en un modèle

décomposable, car sa taille empêche son évaluation complète (simulation) par un humain. Enfin, une augmentation supplémentaire de sa taille et l'utilisation de relations de caractéristiques complexes feront perdre au modèle sa transparence algorithmique, abandonnant les caractéristiques précédentes (Alejandro.B et al, 2019).

4.1.3. K-plus proches voisins

Une autre méthode qui relève des modèles transparents est celle des K-plus proches voisins (KNN), qui traite des problèmes de classification de manière méthodologiquement simple. Elle prédit la classe d'un échantillon de test en votant pour les classes de ses K plus proches voisins (où la relation de voisinage est induite par une mesure de distance entre les échantillons). Lorsqu'elle est utilisée dans le contexte des problèmes de régression, le vote est remplacé par une agrégation (par exemple, la moyenne) des valeurs cibles associées aux plus proches voisins.

En termes d'explicabilité du modèle, il est important d'observer que les prédictions générées par les modèles KNN reposent sur la notion de distance et de similarité entre les exemples, qui peuvent être adaptées en fonction du problème spécifique à traiter. Fait intéressant, cette approche de prédiction ressemble à celle de la prise de décision humaine basée sur l'expérience, qui se fonde sur le résultat de cas similaires passés. C'est cette similarité qui explique pourquoi le KNN a également été largement adopté dans des contextes où l'interprétabilité du modèle est une exigence.

De plus, en plus d'être simple à expliquer, la capacité d'inspecter les raisons pour lesquelles un nouvel échantillon a été classé dans un groupe, ainsi que d'examiner comment ces prédictions évoluent lorsque le nombre de voisins K est augmenté ou diminué, renforce l'interaction entre les utilisateurs et le modèle.

Il faut garder à l'esprit que, Un K très élevé empêche une simulation complète des performances du modèle par un utilisateur humain. De même, l'utilisation de caractéristiques complexes et/ou de fonctions de distance compliquées nuirait à la décomposabilité du modèle, limitant son interprétabilité uniquement à la transparence de ses opérations algorithmiques (Alejandro.B et al, 2019).

4.1.4. Apprentissage basé sur des règles

L'apprentissage basé sur des règles (Rule-based Learning) fait référence à tout modèle qui génère des règles pour caractériser les données dont il est censé apprendre. Les règles peuvent

prendre la forme de simples règles conditionnelles si-alors ou de combinaisons plus complexes de simples règles pour former leur connaissance. Également connecté à cette famille générale de modèles, les systèmes basés sur des règles floues sont conçus pour un champ d'action plus large, permettant la définition de règles formulées verbalement sur des domaines imprécis. Les systèmes flous améliorent deux axes principaux pertinents pour cet article. Tout d'abord, ils permettent des modèles plus compréhensibles puisqu'ils opèrent en termes linguistiques. Deuxièmement, ils performent mieux que les systèmes de règles classiques dans des contextes avec certains degrés d'incertitude. Les apprenants basés sur des règles sont clairement des modèles transparents qui ont souvent été utilisés pour expliquer des modèles complexes en générant des règles qui expliquent leurs prédictions.

Les apprenants basés sur des règles sont d'excellents modèles en termes d'interprétabilité dans divers domaines. Leur relation naturelle et fluide avec le comportement humain les rend très adaptés pour comprendre et expliquer d'autres modèles. Si un certain seuil de couverture est atteint, une enveloppe de règles peut être considérée comme contenant suffisamment d'informations sur un modèle pour expliquer son comportement à un utilisateur non expert, sans renoncer à la possibilité d'utiliser les règles générées comme un modèle de prédiction autonome (Alejandro.B et al, 2019).

4.1.5. Modèle additifgénéral

En statistiques, un modèle additif généralisé (GAM)(Generalized Additive Model) est un modèle linéaire dans lequel la valeur de la variable à prédire est représentée par l'agrégation de plusieurs fonctions lisses inconnues qui correspondent aux variables prédictives. L'objectif principal d'un GAM est d'identifier ces fonctions lisses afin que leur sortie combinée s'approche étroitement de la variable prédite. Cette structure offre un niveau élevé d'interprétabilité, permettant aux utilisateurs d'évaluer l'importance de chaque variable prédictive et de comprendre comment elle influence la sortie prédite à travers sa fonction lisse associée.

Les chercheurs utilisent souvent les GAM non seulement pour leur précision, mais principalement pour leur capacité à éclairer les relations entre les variables dans le jeu de données. Cet accent mis sur la compréhension des problèmes sous-jacents et des interactions entre les variables fait des GAM un choix de modélisation privilégié dans certaines communautés, même s'ils peuvent présenter des performances inférieures par rapport à des modèles plus complexes. L'accent mis sur l'interprétabilité plutôt que sur la précision prédictive pure souligne la valeur des

GAM dansdes contextes où comprendre les données et leurs relations est primordial (Alejandro.B et al, 2019).

4.1.6. Modèles bayésiens

Un modèle bayésien (Bayesian model) prend généralement la forme d'un modèle graphique acyclique dirigé probabiliste dont les liens représentent les dépendances conditionnelles entre un ensemble de variables. Par exemple, un réseau bayésien pourrait représenter les relations probabilistes entre les maladies et les symptômes. Étant donné les symptômes, le réseau peut être utilisé pour calculer les probabilités de la présence de diverses maladies. À l'instar des GAMs, ces modèles fournissent également une représentation claire des relations entre les caractéristiques et la cible, qui dans ce cas sont données explicitement par les connexions reliant les variables entre elles (Alejandro.B et al, 2019).

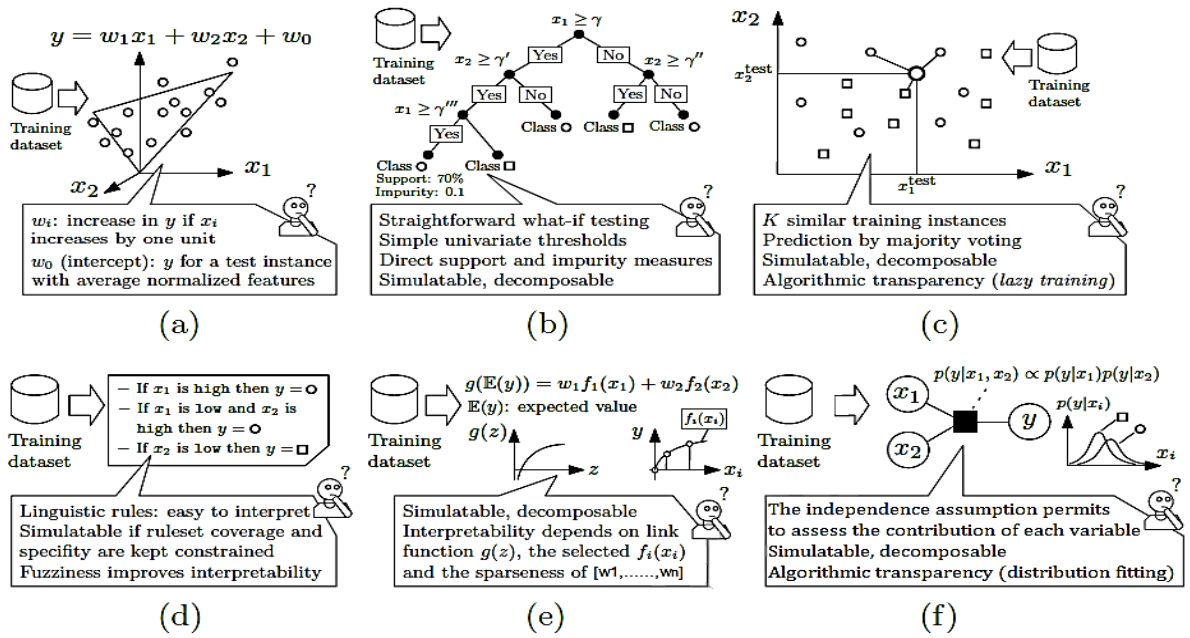

Figure 1.7. Modèles simples explicables : (a) Modèles à linéarité ; (b) Arbres décisionnels ; (c) Apprentissage par instance ; (d) Apprentissage fondé sur des règles ; (e) Modèles à composantes dispersées ; (f) Classificateur Naïve Bayes. (Alejandro.B et al, 2019)

4.2.Les modèles opaques (opaque models)

Les modèles opaques ou bien les algorithmes opaques sont des modèles dont le mécanisme interne est ardu à déchiffrer une fois l'entraînement terminé.Des techniques telles que les forêts aléatoires (RF), les machines à vecteurs de support (SVM), les réseaux de neurones convolutifs

(CNN), les réseaux de neurones multicouches (MNN) et les réseaux de neurones récurrents (RNN)sont souvent privilégiées,malgré leur opacité, car les méthodes plus explicites ne garantissent pas toujours des résultats satisfaisants. Néanmoins, ces modèles sont souvent perçus comme des « boîtes noires » et requièrent donc des démarches d'explicabilité postérieure, destinées à fournir des justifications a posteriori après la fin de l'entraînement.

4.2.1. Forêt Aléatoire

Les forêts aléatoires (RA) ont été initialement proposées pour améliorer la précision des arbres de décision uniques, souvent sujets au surapprentissage et, par conséquent, à une mauvaise généralisation. Les forêts aléatoires corrigent ce problème en combinant plusieurs arbres afin de réduire la variance du modèle résultant et d'améliorer la généralisation. Pour ce faire, chaque arbre est entraîné sur une partie différente du jeu de données d'entraînement, capturant différentes caractéristiques de la distribution des données, afin d'obtenir une prédiction agrégée (Steven.A et al, 2020).

Cette procédure produit des modèles très expressifs et précis, mais au détriment de l'interprétabilité, car la forêt entière est beaucoup plus difficile à expliquer que des arbres uniques, ce qui oblige l'utilisateur à appliquer des techniques d'explicabilité post-hoc pour comprendre le mécanisme de décision (Steven.A et al, 2020).

4.2.2. Machine à Vecteurs de Support

Les SVM forment une classe de modèles profondément ancrés dans les approches géométriques. Initialement introduits pour la classification linéaire, ils ont ensuite été étendus au cas non linéaire, tandis qu'une atténuation du problème initial les a adaptés aux applications réelles. Intuitivement, dans un contexte de classification binaire, les SVM trouvent l'hyperplan de séparation des données avec la marge maximale, ce qui signifie que la distance entre celui-ci et le point de données le plus proche de chaque classe est le plus grand possible. Outre leur utilisation dans la classification, les SVM peuvent être utilisés dans les problèmes de régression, voire de clustering. Bien que les SVM aient été utilisés avec succès dans un large éventail d'applications, leur grande dimensionnalité, leurs transformations de données potentielles et leur motivation géométrique en font des modèles très complexes et opaques (Steven.A et al, 2020).

4.2.3. Réseau de Neurones Multi-couches

Les RN constituent une classe de modèles largement utilisés dans de nombreuses applications, allant de la bio-informatique aux systèmes de recommandation,en raison de leurs performances de pointe. En revanche, leur topologie complexe entrave leur interprétabilité, car on ne sait pas clairement comment les variables interagissent entre elles ni quelles caractéristiques de haut niveau le réseau a pu capter. De plus, même la compréhension théorique et mathématique de leurs propriétés n'est pas suffisamment développée, ce qui en fait de véritables modèles de boîte noire.

D'un point de vue technique, les RN sont constituées de couches successives de nœuds reliant les caractéristiques d'entrée à la variable cible. Chaque nœud d'une couche intermédiaire collecte et agrège les sorties de la couche précédente, puis produit lui-même une sortie en transmettant la valeur agrégée via une fonction (appelée fonction d'activation). Ces valeurs sont ensuite transmises à la couche suivante, et ce processus se poursuit jusqu'à la couche de sortie.

On constate immédiatement que plus le nombre de couches augmente, plus l'interprétation du modèle devient complexe (Steven.A et al, 2020).

4.2.4. Réseau de Neurones Convolutionnel

Les réseaux neuronaux convolutifs, aussi appelés CNN ou ConvNets, sont des modèles de réseau neuronal artificiel à action directe. Il est utilisé en deep Learning pour évaluer les informations visuelles. Ces réseaux sont capables de traiter un grand nombre de tâches impliquant des images, des sons, des textes, des vidéos et d'autres médias (Hallali.H, 2022).

4.2.5. Réseau de Neurones Récurrent

Un réseau neuronal récurrent (RNN) est un type de réseau neuronal artificiel qui utilise des données séquentielles ou des données de séries temporelles. Les réseaux de neurones récurrents utilisent des données de formation pour apprendre. Ils se distinguent par leur "mémoire", car ils utilisent les informations des entrées précédentes pour influencer l'entrée et la sortie actuelles (Hallali.H, 2022).

4.3.Techniques Post-hoc

Ces techniques sont divisées en deux catégories : **les techniques agnostiquesau modèle** pour l'explicabilité post-hoc sont conçues pour être intégrées à n'importe quel modèle dans le but d'extraire des informations de sa procédure de prédiction. Et **les techniques spécifiques au modèle** conçues pour des modèles particuliers :

4.3.1.1. Explication par simplification

On peut dire qu'il s'agit de la technique la plus large dans la catégorie des méthodes post-hoc agnostiques au modèle. Les explications locales sont également présentes dans cette catégorie, car parfois, les modèles simplifiés ne sont représentatifs que de certaines sections d'un modèle. Presque toutes les techniques empruntant cette voie pour la simplification du modèle sont basées sur des techniques d'extraction de règles. Parmi les contributions les plus connues à cette approche (Steven.A et al, 2020) :

- **LIME (Local Interpretable Model-Agnostic Explanations)**

La méthode **LIME** (*Local Interpretable Model-agnostic Explanations*) est largement reconnue comme l’une des techniques les plus utilisées en matière d’explicabilité. Elle se distingue par sa capacité à fournir des explications locales et interprétables, indépendamment de la nature du modèle de prédiction utilisé. L’approche consiste à approximer le comportement d’un modèle complexe dans l’environnement immédiat d’une instance donnée c’est-à-dire une observation spécifique dont on souhaite comprendre la prédiction en s’appuyant sur un modèle simple, tel qu’une régression linéaire ou un arbre de décision.

Pour ce faire, LIME génère plusieurs versions légèrement modifiées de l’instance cible (appelées perturbations) et interroge le modèle complexe pour obtenir les prédictions correspondantes. À partir de ces données, un modèle interprétable est ajusté localement, de manière à reproduire au mieux le comportement du modèle initial dans cette zone restreinte. Ce modèle substitut permet alors d’identifier les caractéristiques les plus influentes dans la prédiction considérée. Grâce à sa simplicité, sa flexibilité et son caractère agnostique, LIME s’est imposée comme une référence pour analyser les décisions des modèles d’apprentissage automatique dans des contextes variés (Steven A. et al., 2020).

- **G-REX**

Une approche initialement introduite dans la programmation génétique, afin d'extraire des règles à partir de données, mais des travaux ultérieurs ont élargi son champ d'application, la rendant capable d'aborder l'explicabilité (Steven.A et al, 2020).

- **CNF (Forme Normale Conjonctive) ou DNF (Forme Normale Disjonctive)**

Certaines approches cherchent à apprendre des règles logiques sous forme normale conjonctive (ET) ou disjonctive (OU), en supposant que les variables d'entrée sont binaires. L'objectif est de construire un modèle de classification capable de reproduire les décisions d'un modèle complexe à l'aide de règles simples et compréhensibles. Ce type de méthode présente l'avantage de générer un ensemble de règles symboliques, naturellement interprétables par l'utilisateur. Ces règles peuvent non seulement servir à expliquer les prédictions du modèle initial, mais aussi être utilisées directement comme modèle prédictif autonome. (Steven.A et al, 2020).

4.3.1.2. Estimation de la pertinence des caractéristiques

Les techniques faisant partie de cette catégorie ont pour objectif de décrire le fonctionnement d'un modèle opaque en hiérarchisant ou en mesurant l'influence, la pertinence ou l'importance de chaque caractéristique dans les prédictions produites par le modèle à expliquer. Cette catégorie regroupe un ensemble de propositions, chacune faisant appel à des approches algorithmiques différentes visant le même objectif (Alejandro.B et al, 2019).

- **SHAP (SHapley Additive exPlanations)**

SHAP (SHapley Additive exPlanations) est une méthode post-hoc, indépendante du modèle, applicable à tout algorithme d'apprentissage automatique. Fondée sur les valeurs de Shapley issues de la théorie des jeux, elle permet d'estimer la contribution de chaque variable (ou caractéristique) à la prédiction finale du modèle.

Pour ce faire, SHAP attribue à chaque variable un score, en analysant toutes les combinaisons possibles de variables (coalitions). En raison de sa complexité computationnelle, une version approximative appelée **KernelSHAP** a été proposée.

SHAP est largement utilisé pour interpréter les modèles, aussi bien localement que globalement. Toutefois, plusieurs limites doivent être prises en compte :

-Dépendance au modèle : les résultats de SHAP varient selon l'algorithme utilisé, ce qui peut produire des classements différents des variables importantes.

-Interprétation des scores : les valeurs SHAP ne représentent pas des poids absolus, mais doivent être interprétées comme un **classement relatif** de l'importance des variables.

-Problème de colinéarité : SHAP suppose l'indépendance entre les variables, ce qui n'est souvent pas le cas. La présence de variables corrélées peut fausser les explications fournies. (Salih.A & al, 2024)

- **Les mesures QII (Quantitative Input Influence)**

Proposées tiennent compte des entrées corrélées, ce qui quantifie l'influence en estimant le changement de performance lorsqu'on utilise l'ensemble de données original par rapport à un ensemble où la caractéristique d'intérêt est remplacée par une quantité aléatoire (Alejandro.B et al, 2019).

En revanche, les auteurs s'appuient sur l'analyse de sensibilité existante (SA) pour construire une analyse de sensibilité globale (Global SA) qui étend l'applicabilité des méthodes existantes (Alejandro.B et al, 2019).

- **ASTRID(Automatic STRucture IDentification method)**

Une méthode qui vise à identifier quels attributs sont utilisés par un classificateur au moment de la prédiction. Ils abordent ce problème en recherchant le plus grand sous-ensemble des caractéristiques originales de sorte que si le modèle est entraîné sur ce sous-ensemble, en omettant le reste des caractéristiques, le modèle résultant aurait une performance aussi bonne que celle du modèle original (Steven.A et al, 2020).

- Enfin, une autre façon de mesurer l'influence d'un point de données sur la décision du modèle provient des diagnostics de suppression. La différence cette fois est que cette approche concerne la mesure de l'influence de l'omission d'un point de données de l'ensemble de données d'entraînement sur la qualité du modèle résultant, ce qui la rend utile pour diverses tâches, telles que le débogage du modèle (Steven.A et al, 2020).

4.3.1.3. Techniques de visualisation

Les techniques d'explication visuelle sont un moyen d'atteindre des explications indépendantes des modèles (Alejandro.B et al, 2019).

- **ICE (Individual Conditional Expectation)**

Représentant la frontière de décision du modèle en fonction d'une seule caractéristique, tandis que les autres restent fixes (Alejandro.B et al, 2019).

- **PD (Partial Dependence)**

Trace la frontière de décision du modèle en fonction d'une seule caractéristique, mais cette fois, les caractéristiques restantes sont moyennées, montrant ainsi l'effet moyen.

Il existe une relation intéressante entre ces deux graphiques, car en moyennant les graphiques ICE de chaque instance d'un ensemble de données, on obtient le graphique PD correspondant (Alejandro.B et al, 2019).

5. Les travaux connexes

A Hybrid Algorithm of ML and XAI to Prevent Breast Cancer: A Strategy to Support Decision Making	SHAP	XGBoost	Des images mammograp-hies Mendeley Data	2023	-Classifier les patientes atteintes ou non de cancer du sein -Identifier les variables significatives influençant le risque de cancer -Fournir des interprétations compréhensibles pour les équipes médicales
Explainable artificial intelligence in breast cancer detection and risk prediction:	SHAP	XGBoost, LightGBM, Gradient	DDSM TCGA	2024	-L'explication des prédictions de modèles pour détecter le cancer du sein -La classification des

A systematic scoping review		Boosting Machines, et Random Forest	GEO MIAS		sous-types de cancer -L'analyse des biomarqueurs cliniques -L'amélioration de l'interprétabilité et de la transparence des modèles d'IA
An Explainable Artificial Intelligence Model for the Classification of Breast Cancer	SHAP Permutation Importance PDP	KNN ANN) XGBoost RF (SVM)	WBC WDBC	2023	- Développement d'un modèle de ML pour classifier les cancers en bénins ou malins - Identification des caractéristiques les plus influentes (e.g., "Bare Nuclei" pour WBC et "Area Worst" pour WDBC). - Utilisation de XAI pour interpréter les décisions des modèles et faciliter leur adoption en milieu clinique
Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review	Grad-CAM SHAP LIME . OMIG	CNN MobileNet-V2. forêts aléatoires XGBoost.	BUSI	2023	- Évaluation des relations entre l'explicabilité et la performance des modèles. - Analyse des défis éthiques et des lacunes dans les recherches actuelles. - Synthèse des

					méthodologies XAI et des outils utilisés.
Exploring Breast Cancer Diagnosis: A Study of SHAP and LIME in XAI-Driven Medical Imaging	SHAP LIME	**CNN**	Breast Ultrasound Images Dataset	2024	- exploration des performances des techniques XAI (SHAP et LIME) pour expliquer les décisions du modèle EfficientNetV2B2 - Évaluation des techniques XAI en termes de précision, rappel, F1-score, et Intersection over Union (IoU) pour leur capacité à identifier et localiser les régions tumorales dans les images d'échographie - Comparaison des performances des techniques SHAP et LIME pour des scénarios où réduire les faux positifs ou les faux négatifs est critique
Explainable Artificial Intelligence Model for Mammogram Breast Cancer Classifiers	RISE LIME	CNN	DDSM INbreast	2023	- Développement d'un modèle d'IA explicable pour classifier les lésions mammaires - Application des techniques XAI pour fournir des explications visuelles qui aident à l'interprétation médicale

Breast Cancer Diagnosis: A Comprehensive Exploration of Explainable Artificial Intelligence (XAI) Techniques	SHAP LIME Grad-CAM CAM	RF XGBoost SVM ResNet EfficientNet VGG DenseNet	DDSM WDBC GEO TCGA	2024	- L'étude explore l'intégration des techniques de XAI avec les modèles ML/DL pour le diagnostic et la classification du cancer du sein
Explainable Artificial Intelligence Methods for Breast Cancer Recognition	LIME SHAP GradCAM DeepSHAP	CNNs	n'est pas mentionné	2024.	Ils ont exploré les méthodes d'IA explicable pour améliorer la compréhension des prédictions faites par les modèles d'IA utilisés dans la reconnaissance du cancer du sein
Comparison of Explainable Artificial Intelligence Model and Radiologist Review Performances to Detect Breast Cancer in 752 Patients	LIME SHAP	RF XGBoost (X2GAI) SVM K-NN Decision Tree Logistic Regression.	Breast-XD Dataset	2024	Développement d'un modèle explicable pour classifier les lésions mammaires en bénignes ou malignes. Comparaison des résultats du modèle X2GAI avec ceux d’un radiologue expérimenté Visualisation des caractéristiques importantes via SHAP et LIME pour valider les décisions du modèle
Histopathology in focus: a review on		CNN	TCGA-		Ils ont suggéré de développer des approches qui utilisent à

explainable multi-modal approaches for breast cancer diagnosis	LIME SHAP Grad-CAM	GNNs	BRCA BDD:CPTAC-BRCA IMPRESS	2024	la fois des données multimodales et des méthodes explicables pour améliorer la confiance des médecins et des patients
A New Computer-Aided Diagnosis System for Breast Cancer Detection from Thermograms Using Metaheuristic Algorithms and Explainable AI	SHAP	SVM	DMR-IR	2014	Intégration de méthodes XAI (SHAP) pour fournir des explications transparentes des décisions du modèle

Tableau 1.1. Tableau comparatif des travaux récents intégrant des méthodes d'intelligence artificielle explicable (XAI) le diagnostic du cancer du sein.

6. Limites et lacunes des articles étudie

Les articles étudiés soulignent l'importance croissante des méthodes d'intelligence artificielle explicable appliquées au diagnostic et à la classification du cancer du sein. Cependant, elles présentent des lacunes importantes, notamment en ce qui concerne les images mammographiques par exemple le premier article propose un algorithme hybride combinant SHAP et XGBoost pour identifier les facteurs de risque de cancer, mais il se limite à une seule base de données (Mendeley Data) et manque de validations pratiques. Le second article est une revue systématique qui analyse plusieurs techniques XAI (SHAP, LIME, Grad-CAM) et bases de données (DDSM, TCGA), mais elle reste descriptive sans tester directement ces méthodes quant au le troisième article applique SHAP et PDP à des modèles classiques (ANN, SVM) sur des bases structurées comme WBC et WDBC, mais ces bases, simplifiées,

ne reflètent pas la complexité des images mammographiques réelles, nécessitant des tests sur des données bruitées et dans l’article suivent explore Grad-CAM et LIME sur des architectures avancées comme DenseNet, mais n’évalue pas si les explications fournies sont réellement compréhensibles pour les radiologues, un aspect crucial pour des validations pratiques et pour la thèse compare SHAP et LIME pour expliquer des modèles appliqués à des images échographiques, mais ne prend pas en compte les mammographies ni l’impact des explications sur la prise de décision médicale. Enfin, un autre travail combine RISE et LIME pour expliquer des modèles CNN appliqués à DDSM et INbreast, mais n’évalue pas la pertinence des explications pour les médecines, limitant leur utilité en milieu clinique.

7. Conclusion

L'IA, et plus précisément l'apprentissage automatique, transforme radicalement l'analyse des images médicales, surtout en ce qui concerne la détection du cancer du sein. Toutefois, les défis essentiels découlent de la complexité et de l'opacité inhérentes à ces modèles : comment assurer que les décisions soient fiables, compréhensibles et utilisables par les professionnels du secteur médical ?

L'intelligence artificielle explicable (XAI) est considérée comme un outil crucial pour aborder cette question. L'XAI utilise des méthodes sophistiquées comme les cartes de saillance, les modèles explicables et les stratégies post-hoc telles que LIME et SHAP pour promouvoir une transparence augmentée. Ceci contribue à renforcer la confiance et l'engagement des spécialistes médicaux envers ces technologies.

L'application de l'IA en médecine est toujours confrontée à des obstacles liés au besoin d'interprétations sur mesure, à la standardisation des techniques XAI et à leur validation dans le cadre clinique. L'association entre performance et explicabilité sera déterminante pour sa réussite.

Dans ce chapitre nous avons examiné les diverses stratégies d'explicabilité mises en œuvre sur les modèles d'apprentissage automatique, en soulignant leurs avantages, leurs contraintes et les possibilités de progression pour une intégration optimale dans le milieu clinique.

Chapitre

II

Cancer du sein

1. Introduction

Le cancer du sein est une pathologie majeure qui touche des millions de femmes à travers le monde.Il s'agit du cancer le plus fréquent chez la femme et représente un enjeu de santé publique prioritaire. Cette maladie complexe résulte d'une prolifération incontrôlée des cellules mammaires, influencée par divers facteurs génétiques, hormonaux et environnementaux. Grâce aux avancées scientifiques, notamment en imagerie médicale et en intelligence artificielle, les stratégies de dépistage et de diagnostic ont considérablement évolué, améliorant ainsi la prise en charge des patientes.Ce chapitre explore en profondeur l'anatomie et la physiopathologie du sein, l'épidémiologie du cancer mammaire ainsi que les principales méthodes utilisées pour sa détection et son diagnostic.

2. Anatomie dusein

Le sein est composé d'une glande mammaire, de fibres de soutien (ligaments de Cooper) et de graisse (tissu adipeux), le tout est recouvert par la peau. La quantité de chacune de ses composantes peut varier d'une femme à l'autre. Le sein est situé par-dessus le muscle pectoral. On trouve également dans le sein des nerfs, des vaisseaux sanguins et lymphatiques. La glande mammaire est divisée en 15 à 20 sections qu'on appelle lobes, composés de lobulés. Ceux-ci sont reliés à des canaux qui se rendent sous le mamelon (situé au centre du sein). On peut également observer des chaînes de ganglions lymphatiques qui filtrent les microbes et protègent le corps contre l'infection et la maladie. Le cancer du sein peut se développer tant au niveau d'un canal galactophore que d'un lobule et il peut également se retrouver au niveau des ganglions lymphatiques (Figure 2.1). (Abderrezak.M et al, 2020)

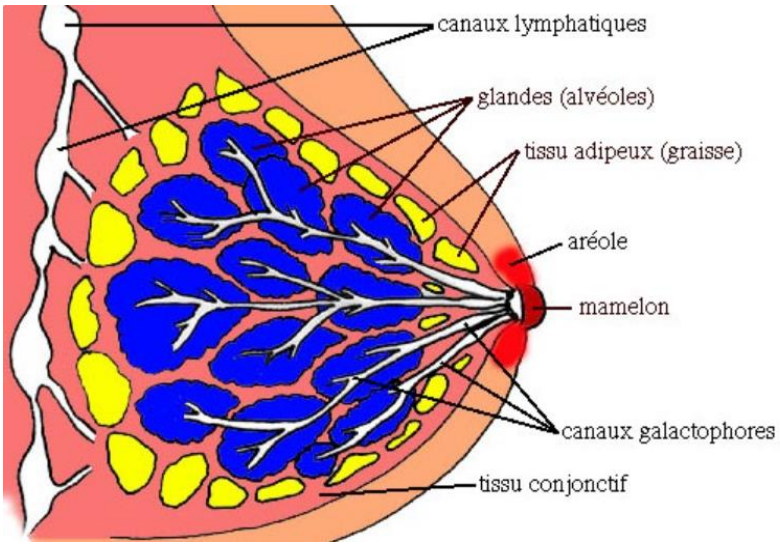

Figure2.1.L'anatomie du sein. (Abderrezak.M et al, 2020)

3. Le cancer du sein

Le cancer du sein est le cancer le plus fréquent chez la femme et la deuxième cause de mortalité parmi tous les cancers. Il se développe à partir des cellules initialement normales qui constituent la glande mammaire. Ces cellules se transforment et se multiplient de façon anarchique et excessive par des mutations ou des instabilités génétique (anomalie cytogénétique) pour former une masse cellulaire appelée tumeur maligne. **(Radia.N et al, 2022)**

3.1. Les différents types de cancer du sien

3.1.1. Cancers du sein suivant leur localisation

Le sein est composé de nombreux lobes glandulaires (eux-mêmes constitués de plusieurs lobules), structures de production du lait. Chacun se poursuit par un canal lactifère, qui l'amène jusqu'au mamelon. Les lobes glandulaires sont entourés de tissu adipeux (graisse), ainsi que des vaisseaux sanguins et lymphatiques.

Les vaisseaux lymphatiques conduisent la lymphe au niveau des ganglions axillaires, situés sous le bras. Ces ganglions, sortes de réservoirs de cellules immunitaires, peuvent parfois être atteints par les cellules tumorales.

Il existe plusieurs cancers du sein, différant selon leur localisation et leur extension. Les trois formes les plus fréquemment rencontrées sont les suivantes **((FRM), 2023)**:

- **Les carcinomes in situ**

Les cellules cancéreuses restent dans les canaux (cancer canalaire in situ) et les lobules (cancer lobulaire in situ), et n'ont pas diffusé dans les tissus environnants. Le cancer canalaire in situ est la forme de cancer du sein in situ la plus fréquente, selon l'INCa **((FRM), 2023)**.

- **Les carcinomes infiltrants**

Les cellules cancéreuses ont envahi les tissus entourant les canaux et lobules. Si ce type de cancer n'est pas pris en charge à temps, il conduit à la formation de métastases dans les ganglions axillaires et le reste du corps **((FRM), 2023)**.

- **Les carcinomes inflammatoires**

Ces cancers se situent en surface, au niveau de la peau.

3.1.2. Cancers du sein selon leurs caractéristiques moléculaires

La recherche sur le génome a permis d'établir une classification plus fine des cancers du sein, basée sur la présence de certains marqueurs dans les tumeurs **((FRM), 2023)**.

Ainsi, on y recherche les protéines : RE (récepteur aux œstrogènes), RP (récepteur à la progestérone), HER2 (récepteur du facteur de croissance épidermique 2) qui permettent de différencier plusieurs groupes de cancers du sein **((FRM), 2023)**.

Les cancers du sein « hormonodépendants » (RH+) encore appelés luminaux sont les formes les plus fréquentes, et sont positifs pour les récepteurs hormonaux aux œstrogènes (RE+) et à la progestérone (RP+). Ceux de type « HER2+ » surexpriment la protéine HER2. Les cancers du sein négatifs pour ces trois marqueurs sont appelés « triple négatifs » **((FRM), 2023)**.

4. Epidémiologie

Globalement, le cancer du sein est le cancer le plus fréquent chez les femmes, tant en termes de mortalité qu'en termes d'incidence, à la fois dans les pays développés et en développement. Il représente également à lui seul 50% des cancers gynécologiques (sein, ovaires, corps et col de l'utérus) (Figure 2.2). **(Sarah.B et al, 2018)**

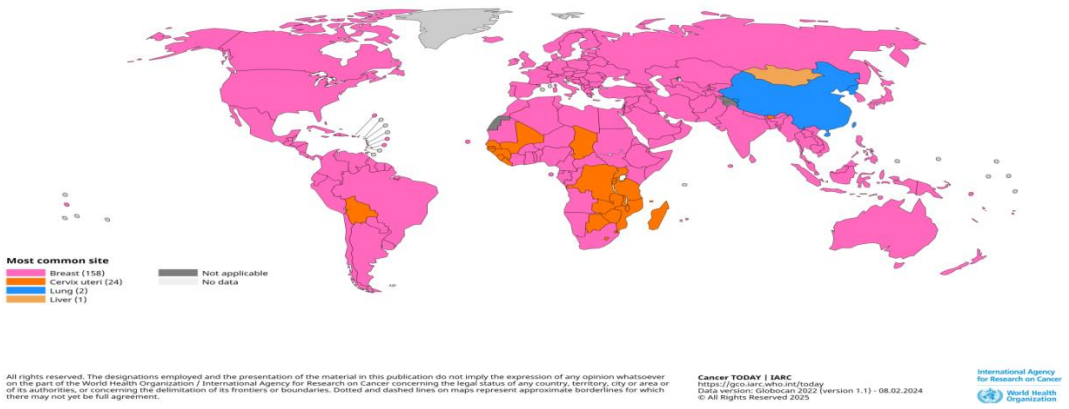

Figure 2.2. Carte mondiale illustre le type de cancer le plus fréquent dans chaque pays, selon les données de Globocan 2022 (Version 1.1) ((CIRC), 2022)

4.1. Epidémiologie mondiale

D'après les résultats annoncés en 2022, on a recensé 2 296 840de cas féminins dans le monde soit un pourcentage égal à 23.8 % et un taux de mortalité par cette maladie égale à 15.4% soit 666 103cas. Présent dans tous les pays, le cancer du sein touche les femmes de tous âges à partir de la puberté, mais son incidence croît à mesure que l'âge avance **((CIRC), 2022)**.

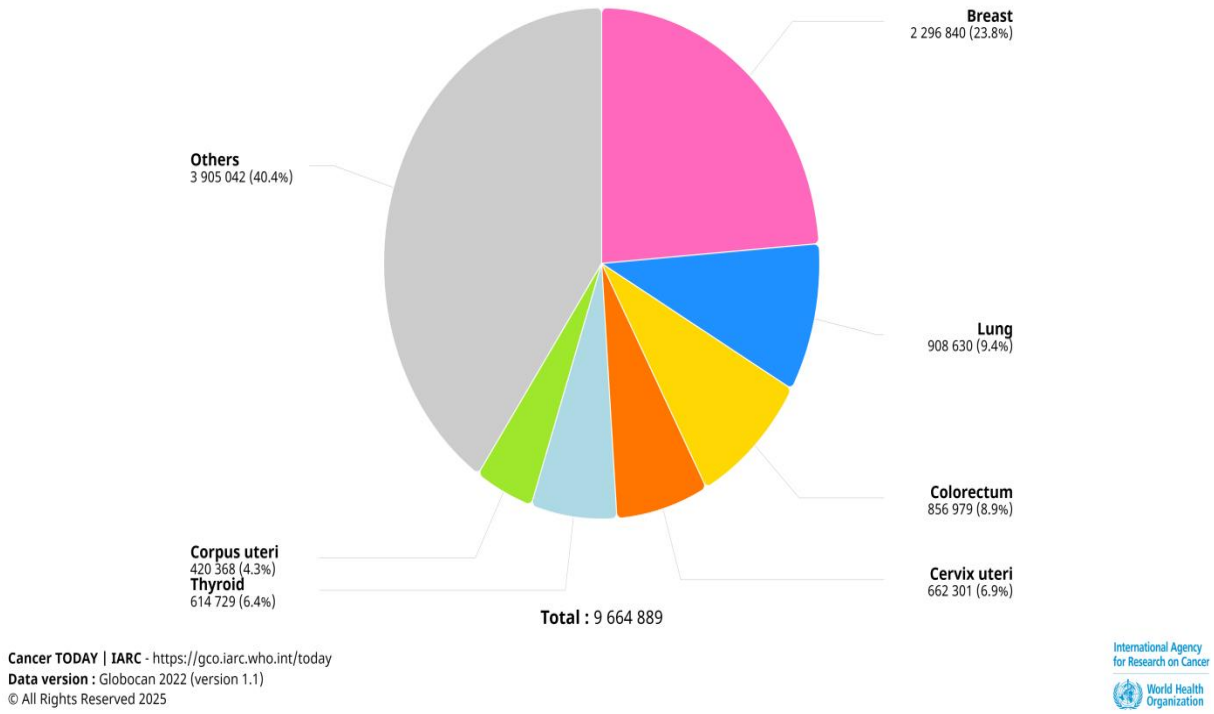

Figure 2.3. Un diagramme circulaire représentant la répartition des cas de cancer dans le monde selon Globocan 2022 (Version 1.1) ((CIRC), 2022)

4.2. Epidémiologie en Afrique

En 2022, l'Afrique a enregistré 198 553 cas de cancer du sein chez les femmes, faisant de cette maladie la principale cause de cancer féminin sur le continent. Le taux de mortalité associé atteint 85 787 décès, révélant une situation particulièrement préoccupante ((CIRC), 2022).

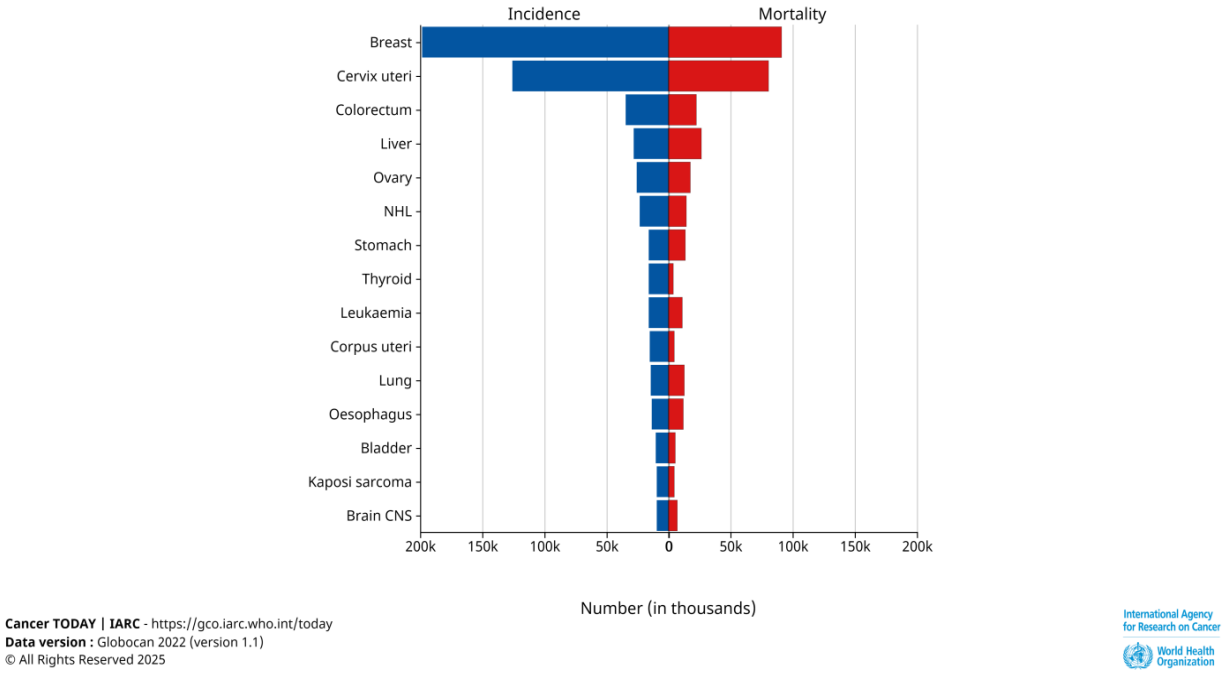

Figure 2.4. Un histogramme représentant l'incidence et la mortalité des cancers féminins en Afrique, selon les données de Globocan 2022 (Version 1.1) ((CIRC), 2022)

4.3. Epidémiologie en Algérie

Le pourcentage total des cas en Algérie est 22,6% pour les deux sexes mais pour les femmes seulement, il est de 41.3% **((CIRC), 2022)**.

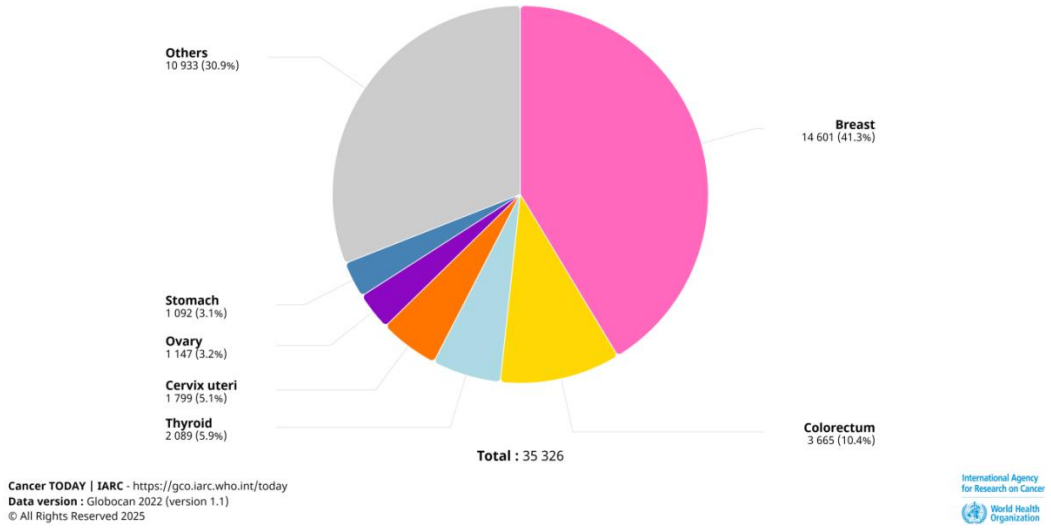

Figure 2.5. Un graphique circulaire représente l'incidence des cancers féminins en Algérie en 2022, selon les données de Globocan 2022 (Version 1.1) ((CIRC), 2022)

5. Facteurs de risque du cancer du sein

Bien que certains facteurs de risque du cancer du sein soient connus, ce cancer est considéré comme une maladie multifactorielle parce qu'il reste plusieurs facteurs inconnus et les causes exacte du cancer du sein ne sont pas bien élucidés **(Radia.N et al, 2022)**.

5.1. L'âge et sexe

L'âge est le facteur de risque le plus important du cancer du sein, l'âge moyen de survenue de cancer du sein est de 55 ans. Le cancer du sein est beaucoup plus agressif chez une femme jeune que chez une femme ménopausée. C'est un cancer quasi exclusif de la femme. Il est 100 fois moins fréquent chez l'homme **(Radia.N et al, 2022)**.

5.2. Les facteurs hormonaux endogènes

- **La puberté et la ménopause**

La survenue des premières règles avant l'âge de 12 ans augmente le risque de cancer du sein, aussi les femmes qui ont leur ménopause après 55 ans **(Radia.N et al, 2022)**.

- **La nulliparité**

Parmi les premiers facteurs qui favoriseraient la survenu du cancer du sein, la nulliparité, ce risque a été prouvé par une comparaison des femmes qui n'ont pas eu d'enfant et les femmes qui ont eu au moins une grossesse chez qui le risque de cancer est réduit de 25% **(Radia.N et al, 2022)**.

Plus l'âge de la première grossesse est jeune plus la protection est grande, dans ce cas la différenciation des glandes mammaires est ensuite moins sensible à l'effet de divers carcinogenèse et l'inverse, plus la grossesse est tardive plus le risque de cancer est supérieur que celui des femmes nullipares **(Radia.N et al, 2022)**.

- **L'allaitement**

Les dernières études montrent que l'allaitement a un rôle protecteur contre le cancer du sein (diminution de 4% pour chaque année d'allaitement), cette protection est due à la sécrétion de

la prolactine dans la période anovulatoire avec une sécrétion des œstrogènes. **(Radia.N et al, 2022)**.

5.3. Les facteurs exogènes

- **Contraceptifs oraux**

Malgré les recherches menées sur ce sujet depuis 60 ans, aucune affirmation sur la relation entre le cancer et les contraceptifs n'a été établie. Certaines études retrouvent un risque de l'ordre de 1.5 chez les femmes jeunes ayant utilisé des CO pendant 5 ans. **(Radia.N et al, 2022)**

- **Traitement hormonal substitutifs (THS)**

Un grand nombre de chercheurs ont estimé que le traitement hormonal substitutif de la ménopause, s'il est associé avec un œstro-progestatif, augmente le risque de cancer du sein post -ménopausique. Le risque est lié à la durée d'utilisation du traitement. **(Radia.N et al, 2022)**

5.4. Les caractéristiques staturo-pondérales, la nutrition et la sédentarité

- **Le poids :** plus le poids est élevé plus le risque d'atteinte par le cancer du sein est élevé aussi **(Radia.N et al, 2022)**.
- **Une alimentation riche en graisse et la consommation d'alcool (Radia.N et al, 2022)** .
- **l'activité physique :** les femmes qui pratiquent une activité physique régulière ont un risque réduit par rapport aux femmes sédentaires **(Radia.N et al, 2022)**

6. Signes cliniques évocateurs de cancer du sein

6.1. Modification de la forme et de l'aspect du sein

Le signe le plus fréquent et facile à reconnaitre lors d'un cancer du sein est la présence d'une boule (ou masse) dans le sein. Généralement, elle n'est pas douloureuse, plutôt dure et ses contours ne sont pas très bien définis. On la découvre souvent de façon fortuite, en palpant le

sein, soit par autopalpation, soit lors d'un examen gynécologique. Cependant, il peut aussi y avoir d'autres modifications du sein qui sont observables sans avoir besoins de palpations. On peut observer un changement de taille, de forme ou d'apparence du sein. Concernant la modification de la peau, on retrouve un changement de couleur tendant sur le rouge, une peau qui apparait anormalement chaude, l'apparition de fossettes, une veine qui grossit et qui devient très apparente, des lésions, des bosses ou même un aspect de peau d'orange. Concernant le mamelon ou l'aréole, on peut observer une croûte, un liquide inhabituel qui s'écoule, un mamelon enfoncé ou encore un changement de coloration **(Lasnier.A, 2024)**.

6.2. Masse au niveau axillaire

Unautre signe fréquemment observé est une masse dure sous le bras au niveau de l'aisselle, là où se trouvent les ganglions axillaires. Cependant, la palpation des ganglions reste indolore**. (Lasnier.A, 2024)**

6.3. Autres signes cliniques

D'autres signes cliniques peuvent apparaître. Par exemple, des douleurs osseuses peuvent être ressenties ainsi que des nausées, des maux de têtes ou encore une vision double. Une fatigue intense et prolongée ou une faiblesse musculaire peuvent également être synonymes de cancer du sein. Une perte d'appétit, une perte de poids où une jaunisse sont également des signes évocateurs. Sur le plan respiratoire, on retrouve l'essoufflement ou une toux qui est souvent due à l'accumulation de liquide autour des poumons. **(Lasnier.A, 2024)**

7. Le diagnostic du cancer du sein

Le corps de la femme est en constante évolution. Parfois, des changements au niveau des seins, qui semblent normaux peuvent être des signes de cancer. C'est dans ce cas qu'intervient le diagnostic. En effet, un diagnostic précoce d'un cancer du sein pourrait augmenter le taux de survie des sujets atteints. Quand une anomalie mammaire est découverte de manière fortuite ou bien lors d'un examen de dépistage, divers examens doivent être effectués pour confirmer ou infirmer le diagnostic : **(Sarah.B et al, 2018)**

7.1. Examen d'imagerie

Plusieurs méthodes de diagnostic sont actuellement utilisées fréquemment dans l'industrie médicale pour étudier le corps humain. Chaque méthode a des utilisations pour divers organes et est sensible à un type spécifique de contraste. De plus, différentes approches peuvent offrir des connaissances complémentaires sur le même tissu. L'imagerie par ultrasons, l'imagerie par résonance magnétique (IRM) et la mammographie sont les techniques d'imagerie médicale utilisées pour la détection et le traitement du cancer du sein. (Imagerie par rayons X). Les différentes méthodes actuellement utilisées, ainsi que leurs caractéristiques, sont présentées dans l'écriture qui suit **(Benmaamar,O, 2023)**.

7.1.1. Échographie

L'idée de base derrière l'échographie est de placer une sonde sur l'épiderme au-dessus de la zone cible. Les vibrations ultrasonores émises par cette sonde traversent les tissus et sont renvoyées à la sonde. Une fois ce signal recueilli, il est traité par un système informatique qui envoie alors une image en direct sur un écran vidéo. La Figure 2.6 montre deux exemplesd'images d'échographie mammaire, l'une avec une tumeur cancéreuse et l'autre avec une lésion bénigne **(Benmaamar,O, 2023)**.

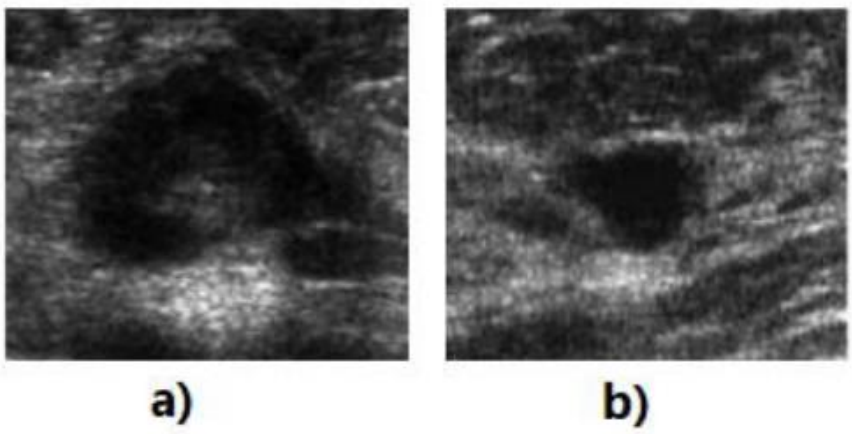

Figure 2.6. Exemples d'images d'échographie mammaire. a) Lésion maligne (sein droit) et b) Lésion bénigne. (Benmaamar,O, 2023)

7.1.2. Imagerie par résonance magnétique (IRM)

Une méthode de dépistage médical relativement nouvelle est l'imagerie par résonance magnétique. (Début des années 1980). Cette technique repose sur l'utilisation d'un rayonnement RF et d'un aimant pour créer le champ magnétique. Les atomes d'hydrogène dans l'organisme humain vibrent subtilement comme principe de base. Tous les atomes d'hydrogène s'alignent de la même manière en présence d'un puissant champ magnétique. Puis, pendant un temps très bref, des signaux radio les stimulent. Ils sont prétendument mis en résonance. Les atomes déchargent l'énergie accumulée en générant un signal à l'issue de cette stimulation. Un dispositif informatique enregistre et convertit ces informations en une image **(Benmaamar,O, 2023)**.

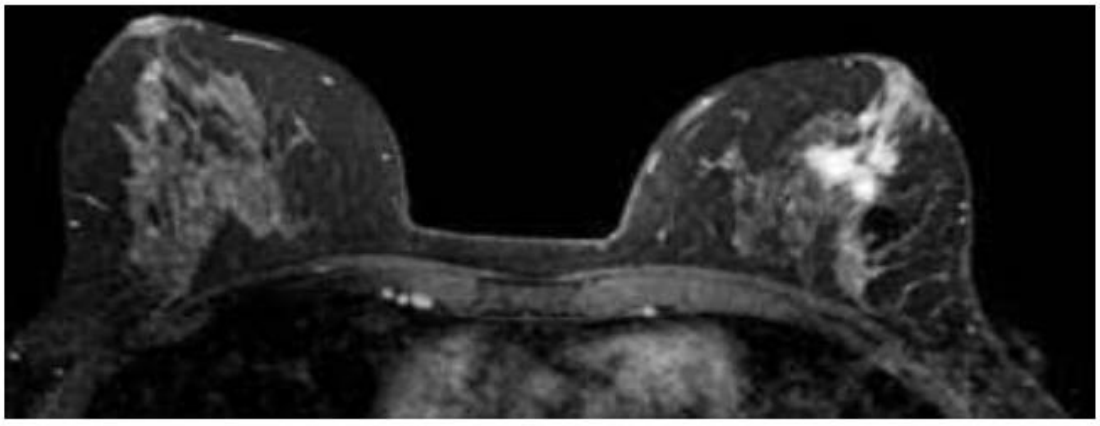

Figure 2.7. Exemple d'IRM mammaire.

L'IRM permet la collecte d'images et d'études de haute qualité dans toutes les dimensions spatiales. De plus, il offre une image haute résolution qui permet une analyse approfondie de la maladie. Cette méthode n'est utilisée que sur des patients sans prothèses métalliques en raison de son coût élevé. Par conséquent, il ne peut être utilisé que pour fournir des informations sur une anomalie déjà identifiée par mammographie ou échographie **(Benmaamar,O, 2023)**.

La mammographie apparaît comme la technique la plus appropriée pour une utilisation générale dans le suivi ou le diagnostic du cancer du sein compte tenu des circonstances générales des différentes méthodes d'imagerie médicale (limites de l'examen échographique, coût de l'examen IRM) **(Benmaamar,O, 2023)**.

7.1.3. Mammographie diagnostique

La mammographie est une technique de radiographie particulièrement adaptée pour examiner les seins des femmes. Elle a pour objectif de détecter les anomalies le plus tôt possible avant qu'elles ne provoquent des symptômes visibles. La mammographie est utilisée non seulement dans les campagnes de dépistage du cancer du sein, mais également pour le diagnostic et la localisation des lésions lors d'interventions chirurgicales telles que les ponctions. Cette technique permet d'observer l'ensemble du tissu mammaire à partir d'une ou deux incidences seulement **(Benmaamar,O, 2023)**.

La mammographie est l'appareil utilisé pour réaliser une mammographie (Figure 8), Il est équipé d'un tube radiogène produisant des rayons X de faible énergie (entre 20 et 50 keV) et d'un système de compression du sein. Les deux seins sont compressés successivement, ce qui permet d'étaler les tissus mammaires et de faciliter la visualisation des structures internes tout en réduisant la dose de rayonnement. Ensuite, les deux seins sont exposés à une faible dose de rayons X, produisant une image sur un détecteur plan. Les images de la glande mammaire sont analysées en distinguant l'atténuation des rayons X par les différents types de tissus. Dans la section suivante, nous détaillons l'anatomie du sein pour établir la relation entre la nature du tissu mammaire et la pénétration des rayons X **(Benmaamar,O, 2023)**.

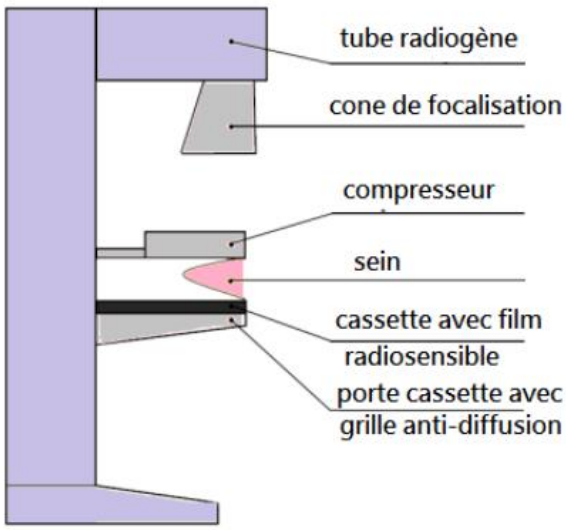

Figure 2.8. Les composants d'une mammographie. (Benmaamar,O, 2023)

7.2. Prélèvement et examen anatomopathologique

La confirmation du diagnostic ne peut être possible que par un examen histologique. Ce dernier est réalisé sur des tissus prélevés par ponction cytologique (cytoponction), qui est une

méthode utilisant une aiguille très fine pour aspirer du liquide ou des cellules provenant de la lésion supposée pathologique. Le prélèvement peut également être récupérer par une biopsie, qui est l'examen permettant de confirmer définitivement si la lésion suspecte est effectivement cancéreuse. Il existe principalement deux types de biopsies : d'abord il y'a la biopsie à l'aiguille, guidée par une mammographie ou une échographie, elle est réalisée en utilisant une fine aiguille qui traverse la peau du sein et prélève un échantillon du tissu anormal. Selon le diamètre de l'aiguille on distingue la macrobipsie (5 à 10 mm) et la microbiopsie (3 à 5 mm) **(Sarah.B et al, 2018)**.

Ensuite, la biopsie chirurgicale permet après une petite incision, d'enlever toute ou partie de la masse suspecte **(Sarah.B et al, 2018)**.

Une fois la biopsie réalisée, l'anatomopathologiste examine, après avoir été traités, les prélèvements tissulaires ou les échantillons de liquides au microscope, à la recherche de cellules cancéreuses. **(Sarah.B et al, 2018)**

8. Dépistage

Dépister c'est-à-dire découvrir ce qui est caché. Le dépistage du cancer du sein est la recherche par un examen systématique chez une femme asymptomatique d'anomalie traduisant une maladie débutante **(Radia.N et al, 2022)**.

Le dépistage intervient dans la phase préclinique appelée « sojourn time » où la maladie n'est pas encore symptomatique mais détectable, l'objectif de la détection précoce par le dépistage est de prévenir ou de retarder la maladie ou atténuer ces conséquences, aussi, il permet de réduire la gravité et la mortalité. En cas d'anomalie, le dépistage permet de prendre en charge les femmes immédiatement **(Radia.N et al, 2022)**.

- Les critères de participation au dépistage organisé du cancer du sein sont :

  -Les femmes porteuses d'une mutation constitutionnelle (prédisposant au cancer du sein).

  -Les femmes ayant des antécédents personnels ou familiaux ou ayant des facteurs de risque.

-Toutes les femmes âgées de 50 à 74 ans **(Radia.N et al, 2022)**.

- Il existe de nombreux dispositifs permettant un dépistage efficace :

8.1. Sensibilisation au cancer du sein

Le manque de connaissances au sujet du cancer du sein et sur ses symptômes est le principal obstacle au dépistage et au diagnostic précoce de la maladie. Il est donc urgent de mettre en place des programmes d'éducation pour sensibiliser et encourager ainsi les femmes à faire attention et à signaler à leurs médecins tout changement inhabituel au niveau des seins. **(Sarah.B et al, 2018)**

8.2. Auto-examen

Les femmes doivent recevoir une formation par leurs médecins afin de pouvoir s'autoexaminer les seins mensuellement à la maison. Cet examen inclut :

- Une inspection visuelle devant un miroir en essayant de noter tout changement de forme ou de taille, des signes de rougeurs ou des altérations du mamelon et de la peau.

- Une autopalpation des seins ainsi que le creux des aisselles en position debout et couchée à la recherche d'une masse palpable.

- Examen du mamelon en essayant de détecter s'il y a écoulement, avec ou sans pincement. **(Sarah.B et al, 2018)**

8.3. Examen clinique

Un examen clinique effectué par un médecin permet de détecter d'éventuelles anomalies qui auraient échappé à la patiente pendant l'auto-examen ou aux techniques d'imagerie comme des signes d'inflammation par exemple. **(Sarah.B et al, 2018)**

8.4. Mammographie de dépistage

Les techniques d'imageries peuvent également être utilisées pour le dépistage. La mammographie est plus particulièrement considérée comme la méthode la plus efficace pour détecter une masse mammaire avant qu'elle ne puisse être palpable **(Sarah.B et al, 2018)**.

9. Diagnostic automatique du cancer (à l'aide de l'IA)

Le diagnostic automatisé utilisé dans l'analyse des mammographies est considéré comme un domaine stratégique dans lequel les modèles d'intelligence artificielle, notamment les réseaux de neurones convolutifs, jouent un rôle de premier plan. Ces techniques visent à améliorer ou à compléter l'expertise médicale en analysant des images pour détecter des anomalies telles que des tumeurs ou des masses qui peuvent indiquer la présence d'un cancer.

Le diagnostic automatique repose sur l'utilisation de modèles d'IA , par exemple , les CNN ont la capacite d'extraire des caractéristiques complexes des images automatiquement , telles que les textures ,le formes pour effectuer des taches de classification multi-classes ou binaire (cancer – non cancer).Le processus de diagnostic automatique généralement comprend plusieurs étapes :

-Le prétraitement des images (améliorer la qualité des images et normaliser)

-L'extraction automatique de caractéristiques complexes (à partirdes caractéristiques extraites, le modèle de classification attribue un label à l'image, par exemple : cancer – non cancer)

-Après l'affichage du diagnostic.

Pour assurer la fiabilité du diagnostic, il est essentiel d'évaluer la performance et l'efficacité du modèleen utilisantdifférentes métriques comme la précision, la spécificité, et la matrice de confusion, ces mesures permettent de vérifier dans quellemesure les prédictions du modèle correspondent aux labels réels.

10.Conclusion

Le cancer du sein est une maladie courante qui peut toucher toutes les femmes, quel que soit leur âge. Il résulte d'une multiplication anormale des cellules mammaires, influencée par des facteurs génétiques, hormonaux et environnementaux.Ce chapitre explore son fonctionnement, ses différentes formes (carcinomes in situ, infiltrants et inflammatoires), ainsi que les caractéristiques moléculaires qui permettent de mieux cibler les traitements.

L'importance du dépistage précoce est fortement mise en avant, car il permet de détecter la maladie avant l'apparition des symptômes, améliorant ainsi les chances de traitement et de survie. Grâce aux avancées médicales, plusieurs méthodes comme la mammographie, l'échographie et l'IRM permettent un diagnostic plus précis. La prévention, incluant un mode de vie sain et une sensibilisation régulière, reste essentielle pour réduire les risques et améliorer les chances de guérison.

Chapitre

III

Méthodologie

1. Introduction

Nous allons présenter dans ce chapitre le contexte général de notre projet qui se base sur l'application de techniques d'intelligence artificielle pour la détection du cancer du sein à partir d'images mammographiques. Nous allons décrive le jeu de données, les principales étapes du prétraitement des images, ainsi que les méthodes d'IA explicable choisies. Notre objectif est d'évaluer et de comparer les différentes approches d'explicabilité appliquées au modèle de classification binaire (cancer – non cancer).

2. Outils logiciels et plateforme d'exécution

2.1. Langage de programmation

Le développement du notre projet s'est appuyé sur le langage **Python**, qui est un langage de programmation interprété, orienté objet et de haut niveau, doté d'une sémantique dynamique. Ses structures de données intégrées de haut niveau, combinées à un typage et uneliaison dynamique, le rendent particulièrement attractif pourle développement rapide d'applications, ainsi que pour une utilisation comme langage de script ou de liaison pour connecter des composants existants (Van.R et al).

Le choix de Python comme langage de programmation pour ce projet repose sur plusieursfacteurs techniques et pratiques :

- Python a été choisi parce qu'il offre plusieurs avantages clés qui le rendent idéal pour les tâches liées à l'intelligence artificielle et au traitement d'images médicales,il se démarque par la simplicité et la lisibilité de sa syntaxe.Son code est transparent et logique, ce qui le rend non seulement facile à écrire, mais aussi à comprendre et à maintenir, même dans le contexte de projets complexes. Aussi, il bénéficie de bibliothèques dédiées, où l'on trouve : NumPy et Pandas pour le traitement de données, TensorFlow, Keras et PyTorch sont utilisés pour l'apprentissage profond et OpenCV et scikit-image sont utilisés pour le traitement d'images. Matplotlib et Seaborn pour la visualisation des résultats.
- Un autre avantage majeur réside dans la vaste communauté Python et la disponibilité de ressources comme des tutoriels et des forums permettent de résoudre rapidement les problèmes et de progresser efficacement dans le projet.

- Finalement, Python peut être compatible avec plusieurs systèmes d'exploitation (Windows, Linux, macOS) et peut facilement être intégré à d'autres outils et langages. Sa présence répandue dans les domaines académiques et professionnels, notamment en santé et en intelligence artificielle, le rend pertinent pour ce projet axé sur le diagnostic du cancer du sein.

2.2. Environnement de développement

L'environnement de développement choisi pour les expérimentations est la plateforme Kaggle. C'est une plateforme largement reconnue par les passionnés des sciences des données et d'apprentissage automatique. Elle offre un environnement collaboratif pour l'analyse de données, la création de modèles et le partage d'informations (Academy).

Nous avons choisi la plateforme Kagglecomme environnement de développement pour ce projet pourdiverses raisons :

- Plateforme accessible en ligne, ne requérant aucune installation.
- Accès gratuit à des GPU de haute performance (tels que le Tesla P100), adaptés aux modèles de deeplearning.
- Compatible avec Python et les bibliothèques couramment utilisées comme TensorFlow, Keras etPyTorch.
- Intégration facile des jeux de données stockés sur Kaggle.
- Cette plateforme offre la possibilité de concevoir et de mettre en œuvre des notebooks interactifs, similaires à Jupyter, sans quitter le navigateur.
- Une communauté dynamique et riche en ressources éducatives et de projets open source.

2.3. Bibliothèques et frameworks

Notre projet s'appuie sur un ensemble structuré de bibliothèque Python open source, choisies en fonction des besoins spécifiques liés à la classification d'images mammographiques et à l'explicabilité des modèles d'apprentissage profond. Le tableau ci-dessous illustre les bibliothèques principales utilisées :

Apprentissage profond	**os , copy**	Gestion des chemins d'accès et manipulation de fichiers
	scikit-learn (train-test-split, accuracy-score,..)	Séparation des jeux de données , calcul des scores (accuracy , F1, rappel , précision) , matrice de connfusion.
	tqdm	Suivi de la progression de l'entrainement et de l'évaluation
Prétraitement des images	**PIL(image) , OpenCV**	Lecured'image,normalisation, redimensionnement (modification de la taille) et ajustement du contraste
Apprentissage profond	**troch.utils.data DataLoader random_split**	Importation et division des jeux de données
	torch, torch.nn, torch.nn.fuctional	Définition de modèles personnalisés sous PyTorch
	torchinfo.summary	Résumé de l'architecture du modèle
Explicabilité des modèles	**LIME**	Visualisation local des zones de l'image qui ont influencé la prédiction
	SHap	Analyse globale et locale de l'importance des caractéristiques
	Grad-Cam	Génères des cartes de chaleur localisant les zones distinctives dans les images

Tableau 3.1. Tableau récapitulatif des bibliothèques utilisées.

2.4. Organisation et traçabilité des expérimentations

Toutes les expérimentations ont été organisées et structurées à l'aide du notebook **kaggle**, ce qui facilité la traçabilité du processus de développement. Notre projet a été mené en plusieurs phases, d'abord, on importe les bibliothèques nécessaires après on charge et on divise le jeu de données, la phase suivante a concerné le prétraitement des images (redimensionnement, normalisation, …). Ensuite, la phase de construction du modèle et cette

étape a été suivie par l'entrainement et l'évaluation de modèle. Enfin on va appliquer les approches de l'explicabilité pour faire la comparaison.

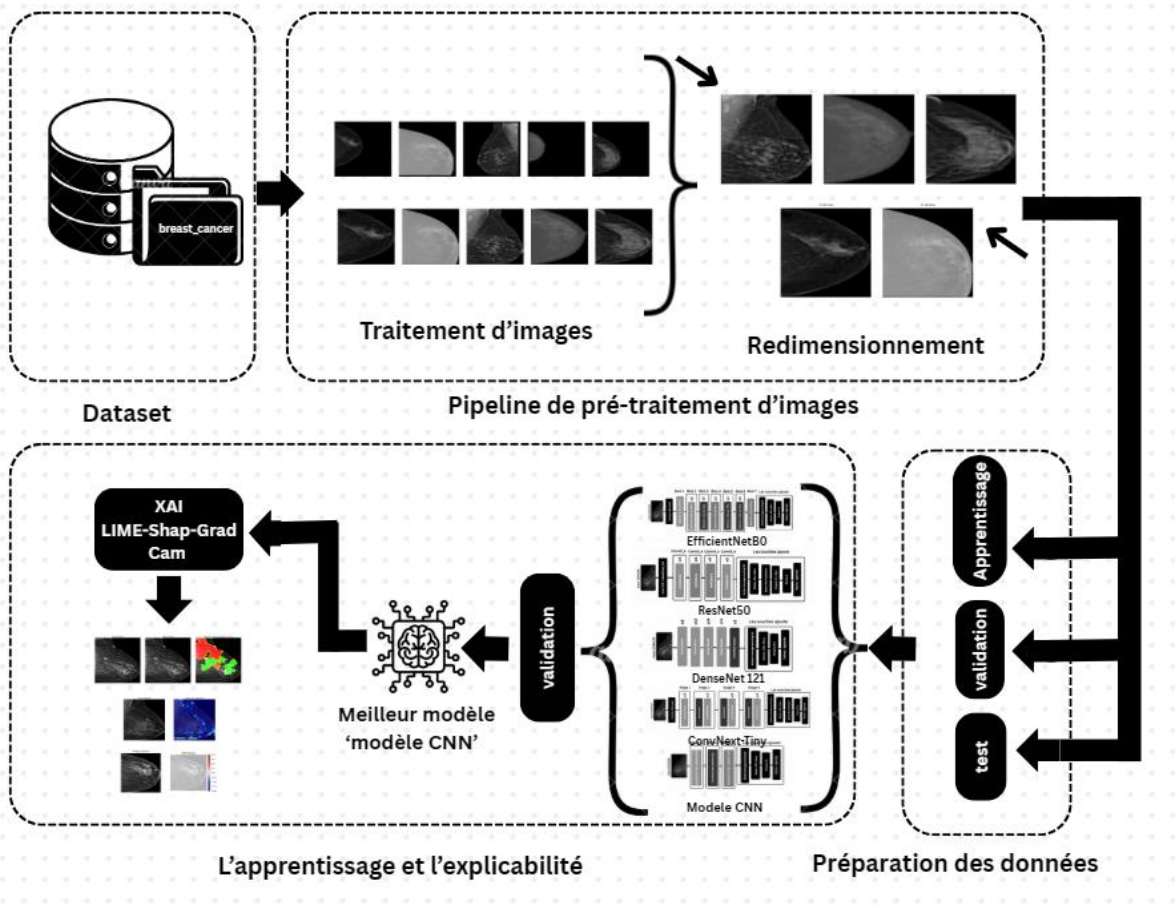

Figure 3.1. Architecture générale de notre système.

3. Données utilisées

3.1. Description de la base de données

Le jeu de données utilisé dans ce projet nommé ***''Final-Final-RSNA-Breast-Cancer-Dataset ''*** est une collection d'images mammographiques pour indiquer la présence ou l'absence de cancer il comprend **487 000 images** de **format png**, diviser en deux parties : **train** (**397 000 images**, les annotations se répartissent comme suit : **183 000 images** de cas de **cancer**,**214 000images non cancer**) et la partie **test** (**90 000 images**,**47 000 images cancer** et **43 000 images non cancer**). Ce jeu de données est accessible sur kaggle :***Final-Final-RSNA-Breast-Cancer-Dataset***

Voici un graphe illustrant la distribution des images dans le dossier **Train**, en fonction des deux classes : **"cancer"** et **"non_cancer"**.

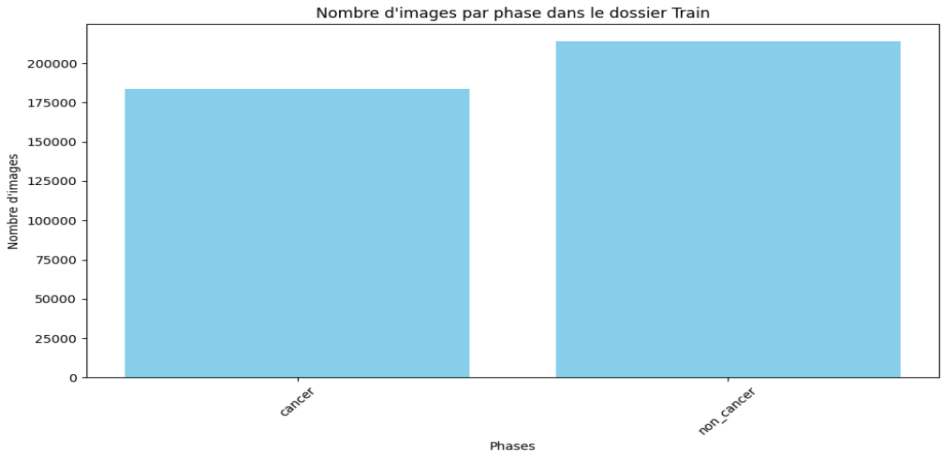

Figure 3.2. Distribution du nombre d'images par phase (la phase d'entrainement).

Voici un graphe illustrant la distribution des images dans le dossier **Test**, en fonction des deux classes : **"cancer"** et **"non_cancer"**:

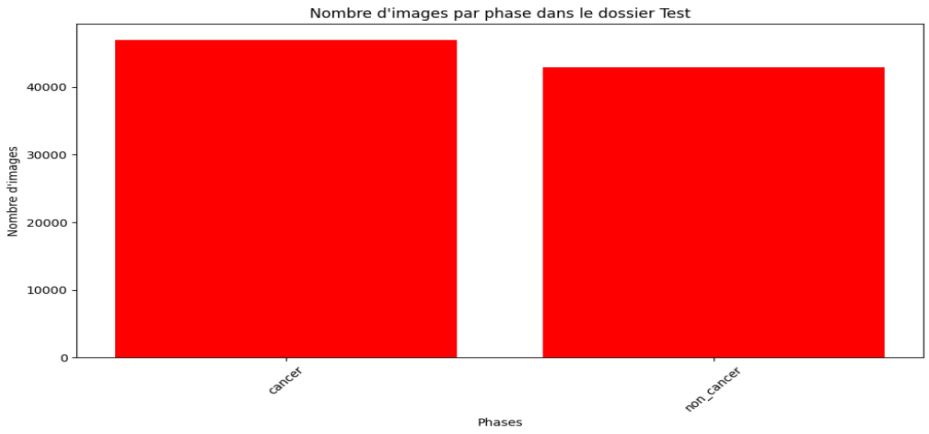

Figure 3.3. Distribution du nombre d'images par phase (la phase de test).

3.2. Séparation des données

Pour assurer une évaluation équitable des performances du modèle, on va diviser le dataset en trois sous-ensembles de cette manière :

-70% des données pour l'entrainement.

-15% des données pour la validation.

-15% des données pour le test.

La division à été réalisée en conservant la distribution des classes (cancer –non cancer) et en garantissant que la même image ne soit pas retrouvée dans tous les sous-ensembles (train , test , val) . Le tableau suivant présente le nombre des images dans chaque section :

Entrainement	278 266
Validation	59 628
Test	59 629

Tableau 3.2. Tableau récapitulatif des données .

3.3. Prétraitement des images

Lors de l'analyse des images mammographies, nous avons remarqué que le fond noir important et l'orientation variables des seins peuvent être problématique. Cet arrière-plan ajoute du bruit inutile et peut distraire les modèles, tandis que les différentes orientations rendent l'apprentissage plus difficile.Pour éviter ces problèmes, on a donc appliqué un prétraitement : on isole la zone du sein par un seuillage et l'identification du contour le plus large, ce qui nous permet d'éliminer le fond noir et après on recadre l'image sur cette zone d'intérêt. Par la suite, on examine la direction du sein en comparant les intensités à gauche et à droite, puis on inverse les images si besoin pour que tous les seins soientorientésdans la même direction.

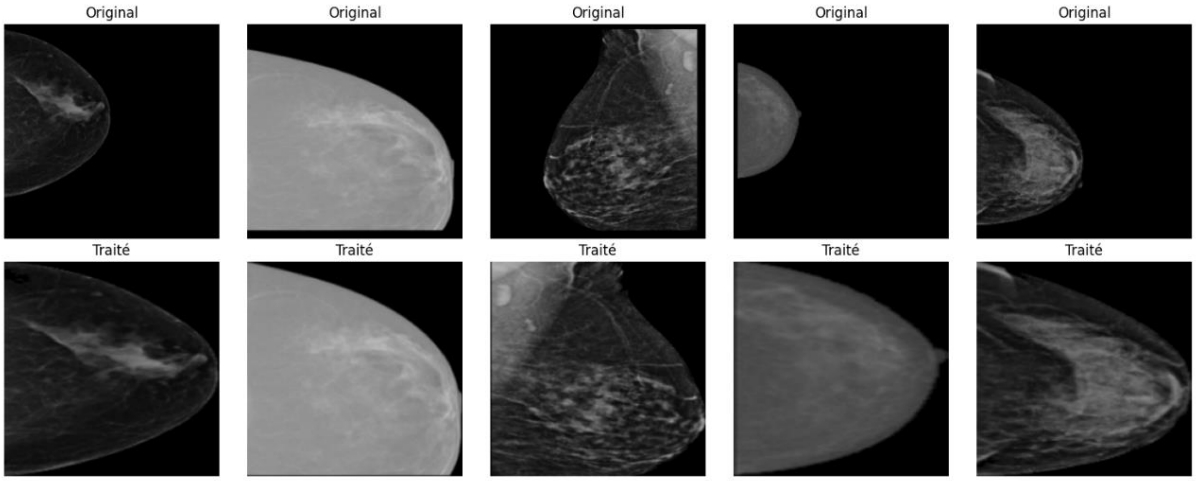

Figure 3.4. Le prétraitement : recadrage sans fond noir et inversion pour orientation uniforme.

Et pour rendre les images homogènes et prêtes à l'utilisation pour l'entrainement, on va appliquer un redimensionnement à une taille fixe 224x224 pixels afin qu'elles aient toutes les mêmes dimensions. Ensuite,une normalisation qui est nécessaire parce qu'elle facilite la convergence et la stabilité du modèle et évite les différences d'échelle des donnéesen convertissant les valeurs de pixels en nombres décimaux compris entre 0 et 1 en divisant par 255.

4. *Choix et construction du modèle*

4.1. EfficientNetB0

La première architecture qu'on a utilisée est **EfficientNetB0**, un réseau neuronal convolutif entrainé sur plus d'un million d'image issues de la base de données **ImagesNet** (MathWorks.efficientnetb0)**,** c'est ce qui le rend capable d'extraire automatiquement des caractéristiques visuelles importantes. Donc on a chargé le modèle sans sa couche de classification finale **(include_top=False)** en réentraînant uniquement les dernières couches du modèle (à partir de couche 100), pour adapter les représentations aux données spécifiques sans perdre les connaissances générales déjà apprises. Puis une couche de résumé (Pooling**'GlobalAveragePooling2D')** pour réduire les dimensions. Ensuite, on ajoute une couche **Dense de 256 neurones** avec l'activation de la fonction **ReLu** pour apprendre les représentations plus complexes**,** et pour limiter le surapprentissage on ajoute un **Dropout** à **50%.** Enfin, notre modèle se termine par une couche de sortie qui utilise une fonction **sigmoïde**pour attribuer une probabilité indiquant si l'image montre un cancer ou non. Il a été entrainé avec l'optimisateur **Adam,** et un taux d'apprentissage (0.0001).

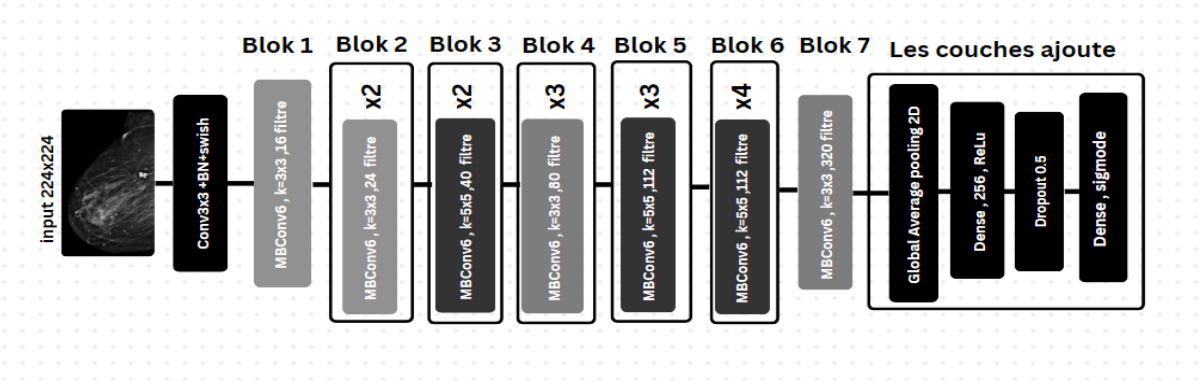

Figure 3.5.L'architecture deEfficientNetB0.

4.2. ResNet50

On a utilisé aussi le modèle**ResNet50,** un réseau de neurones à convolution d'une profondeur de **50 couches ,** qui a été préentrainé sur plus d'un million d'images de la base de données **ImageNet** (MathWorks.resnet50) . On a l'utilisé comme point de départ pour construire un classificateur binaire, et pour adapter à notre tache on a désactivé la couche de sortie d'origine par l'option (**include_top=False**) , ce qui nous a permis de construire notre propre architecture. Tout d'abord on applique un Fine-tuning, en gelant les **80 premières couches** du réseau et en autorisant l'entrainement des couches restant. Par la suite, on construit une

nouvelle tête de classification intégrant une couche **Pooling**(**GlobalAveragePooling2D**), suivi des deux couches **Dense** de **512** et **256 neurones** avec la fonction d'activation **ReLu**. Afin d'améliorer la robustesse du modèle, on ajoute une **régularisation L2**, des **Batch Normalisation** et des **Dropout** (**0,5** et **0,3**respectivement). Finalement, une couche sigmoïde permet de génère une sortie binaire. Notre modèle à été compilé avec l'optimiseur **AdamW**pour optimiser la généralisation parce que il comprend la pénalisation **weight-decay** .

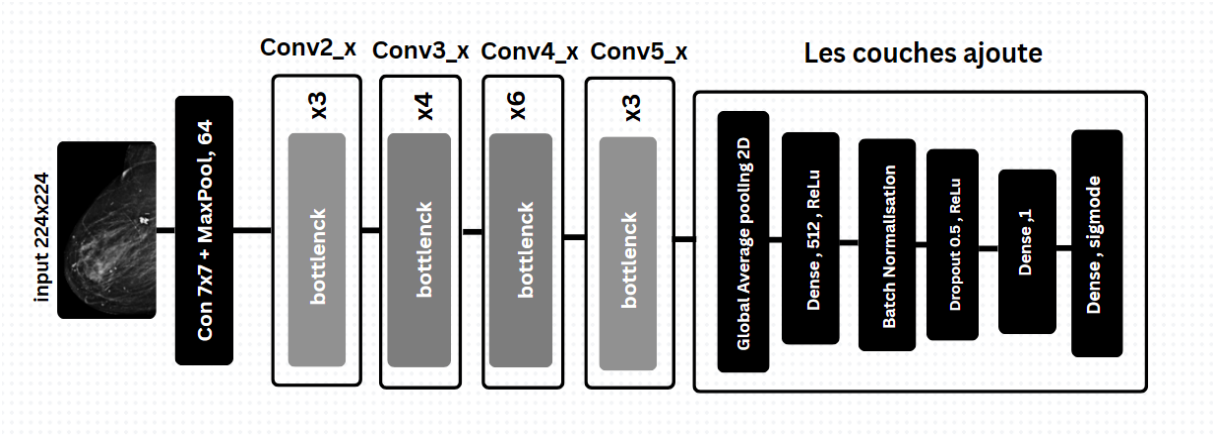

Figure 3.6.L'architecture de**ResNet50** .

4.3. DenseNet121

Lors de notre expérimentation, on a utilisé aussi le modèle **DenseNet121,** qui est un modèle pré-entrainé sur **ImageNet**, comme point de départ pour bénéficier de ses performances en extraction de caractéristiques mais on a retiré la couche de classification d'origine pour pouvoir adapter le modèle à notre phase de classification binaire (cancer – non cancer). Donc on active le Fine-tuning , maisnous gelons les100 premières couches afin de préserver les connaissances générales du modèle .Ensuite, on ajoute une couche **Pooling**(**GlobalAveragePooling2D**) pour réduire la dimension des sorties, après une couche Dense **de 256 neurones** avec la fonction **ReLu**,suivie d'un **Dropout** de **0.5** pour réduire le risque de surapprentissage et comme les modèles précédents on a utilisé la fonction **sigmoïde**dans la dernière couche avec le même optimiseur **Adam**.

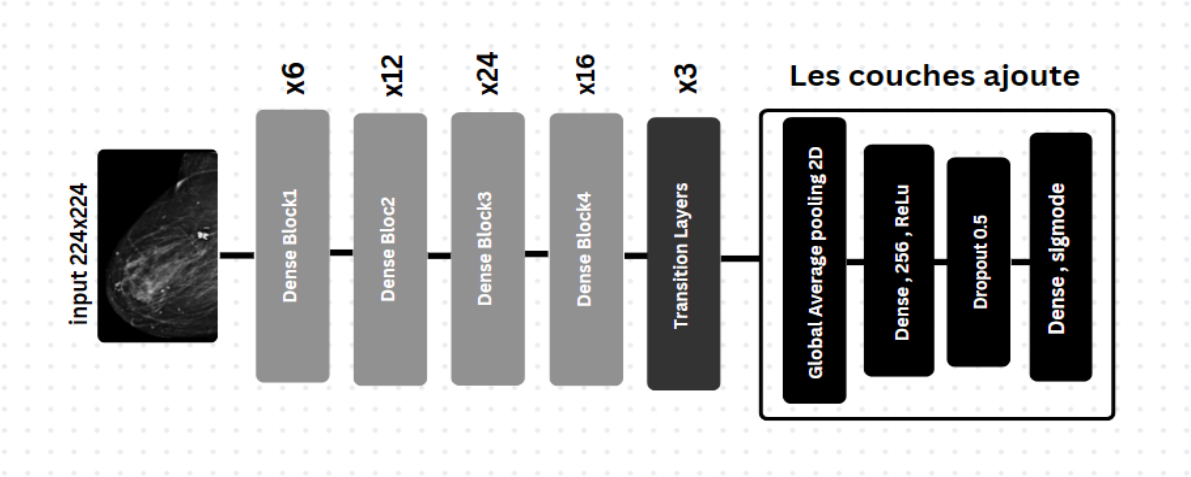

Figure 3.7.L'architecture de **DenseNet 121**.

4.4. ConvNext-Tiny

Comme pour le modèle **DenseNet121**on a utilisé la même stratégie avec ConvNext-**Tiny,** on a chargé le modèle mais on a supprimé la tête de classification afin de pouvoir ajouter nos propres couches, et comme on a déjà fait, en gelant les 100 premières couches (pour conserver les connaissances générales). Puis on ajoute la couche **GlobalAveragePooling2D**, après une couche **Dense de 256 neurones** avec activation **ReLu**et on a intégré un **Dropout**de **0.5**et finalement la couche de sortie (**sigmoïde**). Notre modèle est entrainé avec le même optimisateur **Adam**avec un taux d'apprentissage (**0.0001**).

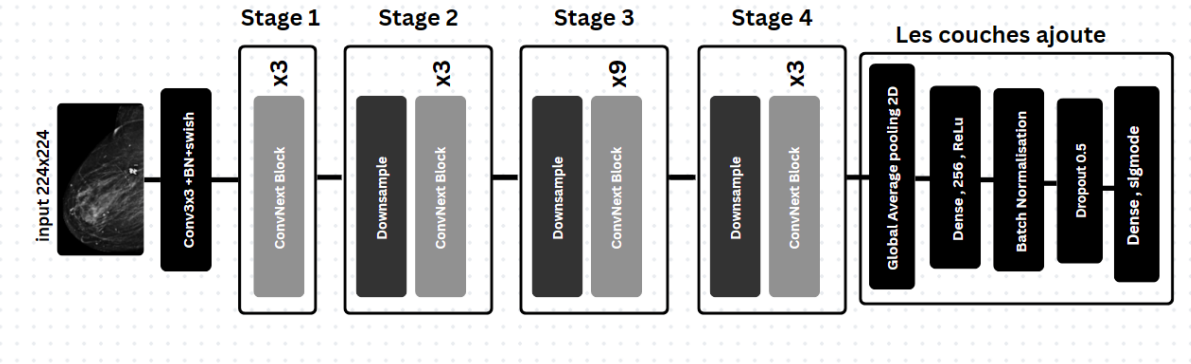

Figure 3.8.L'architecture de ConvNext-Tiny.

4.5. Modele CNN

Nous avons aussi développé un modèle de réseau de neurones convolutifs (CNN) pour la classification binaire d'images mammographiques, dans le but de détecter automatiquement si un cas de cancer ou pas. L'architecture de notre modèle a été construite entièrement sur mesure, sans utiliser de modèle pré-entrainées, ce qui a permis d'adapter l'apprentissage aux données spécifiques. Notre modèle est structuré en plusieurs étapes : Tout d'abord, l'image

traitée par une quatre blocs successifs, chacun comprenant une **couche de convolution** pour détecter les motifs visuels, après une **normalisation**, pour stabiliser le processus d'apprentissage, suivie d'un **Max Pooling**servant à réduire les dimensions tout en préservant les informations importantes. À chaque bloc le nombre de filtre augmente , ce qui offre au réseau la possibilité d'apprendre des représentation visuelles de plus en plus détaillées .Puis , une couche de **Global Average Pooling2D** résume toutes les informations extraites , avant de se diriger vers une couche **Dense** composée de **256 neurones**, et pour en évite le surapprentissage , on a intégré un **Dropout à 50%** .Finalement , pour la sortie on a utilisé une activation **sigmoïde**qui donne une probabilité déterminant si l'image est considérée comme cancéreuse ou non . On a utilisé le même optimisateur **Adam**avec un taux d'apprentissage (0.0005).

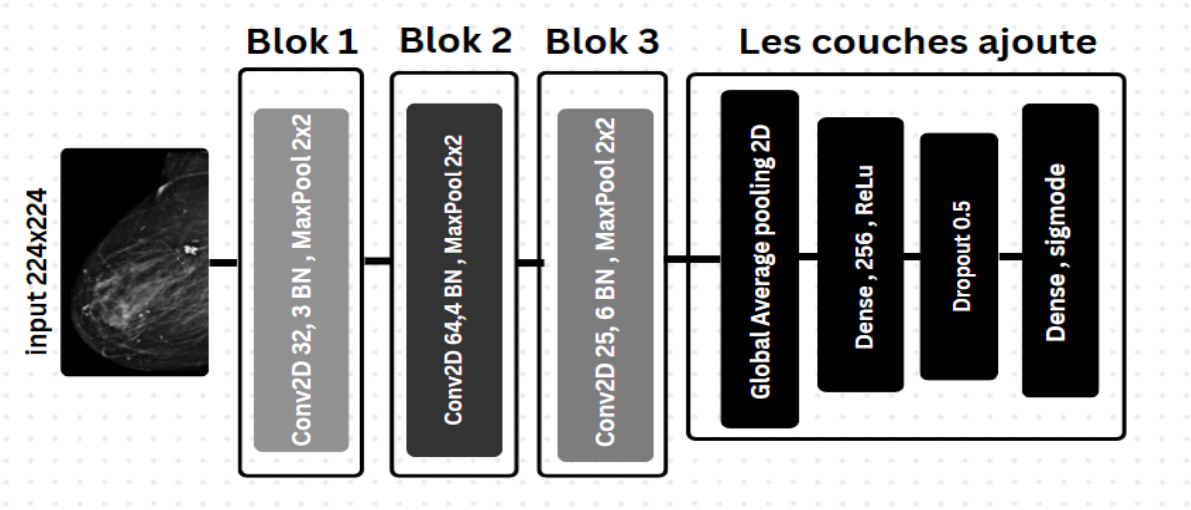

Figure 3.9.L'architecture de CNN .

5. *Les différentes méthodes d'IA explicable*

*5.1.*LIME (*Local Interpretable Model-agnosticExplanations*)

Pour interpréter les décisions prises par notre modèle de classification des mammographies, nous avons eu recours à la méthode **LIME**. LIME est une technique d'explicabilité locale qui consiste à générer des explications compréhensibles pour l'humain en se focalisant sur une instance spécifique (dans notre cas, une image).

Concrètement, nous avons sélectionné une image du jeu de test et avons appliqué la méthode LIME pour identifier visuellement les zones de l'image ayant le plus influencé la prédiction du modèle. Pour cela, nous avons suivi les étapes suivantes :

- **Chargement de l'image**

L'image testée est d'abord convertie en format RGB, puis redimensionnée et transformée de la même manière que lors de l'entraînement du modèle, avant d'être soumise au classifieur pour obtenir une prédiction.

- **Préparation d'une fonction de prédiction pour LIME**

Étant donné que LIME nécessite une fonction prenant en entrée des images au format NumPy et retournant des probabilités, nous avons défini une fonction qui transforme les images, les passe dans le modèle (en mode évaluation), et applique une sigmoïde pour obtenir les probabilités pour chaque classe (cancer / non-cancer).

- **Génération de l'explication avec LIME**

LIME a été utilisé pour perturber l'image originale en masquant certaines régions (appelées *superpixels*) et analyser l'impact de ces perturbations sur la sortie du modèle. À partir de ces observations, LIME a généré une carte visuelle mettant en évidence les zones les plus influentes.

- **Affichage des résultats**

Nous avons affiché trois images côte à côte :
-l'image originale,
-l'image prétraitée,
-l'image annotée avec les résultats de LIME, indiquant les zones importantes pour Les superpixels les plus influents sont surlignés par des contours colorés, permettant une interprétation visuelle intuitive de la décision du modèle.
Grâce à cette approche, nous avons pu vérifier si le modèle s'appuyait sur des régions anatomiquement cohérentes pour établir son diagnostic. Cela constitue une étape importante pour juger de la pertinence clinique du modèle et renforcer la transparence de son fonctionnement.

5.2. SHAP (SHapley Additive exPlanations)

Pour mieux comprendre comment notre modèle prend ses décisions lors de la classification des mammographies, nous avons utilisé la méthode SHAP. Cette technique permet de

visualiser, pour chaque prédiction, quelles zones de l'image ont le plus influencé la décision, en attribuant à chaque région une "valeur d'impact" positive ou négative.
Voici comment nous avons procédé :

- **Préparation de l'image**

Nous avons d'abord sélectionné une image issue du jeu de données, que nous avons convertie en format RGB. Elle a ensuite été redimensionnée à 224×224 pixels et normalisée selon les standards d'ImageNet. Une fois transformée en tenseur, l'image a été ajoutée à un lot (batch) pour qu'elle soit compatible avec le modèle.

- **Prédiction avec le modèle**

L'image prétraitée a été passée dans le modèle en mode évaluation. Comme il s'agit d'un problème binaire, une fonction sigmoïde a été appliquée à la sortie du modèle pour obtenir une probabilité entre 0 et 1. La classe prédite est ensuite déterminée en fonction d'un seuil de 0,5.

- **Génération des explications SHAP**

À l'aide d'un explainer adapté (comme DeepExplainer ou GradientExplainer), nous avons calculé les valeurs SHAP pour l'image. Ces valeurs indiquent dans quelle mesure chaque pixel (ou groupe de pixels) a contribué à la prédiction finale. Cela nous permet de repérer les régions de l'image qui ont le plus pesé dans la décision du modèle.

- **Visualisation des résultats**

Nous avons d'abord affiché l'image en niveaux de gris, puis superposé la carte SHAP avec une échelle de couleurs : les tons rouges indiquent les zones ayant fortement poussé vers la prédiction (par exemple, présence de cancer), tandis que les tons bleus signalent les zones qui ont plutôt joué contre. Une barre de couleur a été ajoutée pour faciliter la lecture, accompagnée d'un petit encart affichant la classe prédite et la probabilité associée.
Cette visualisation nous aide à vérifier si le modèle s'appuie sur des zones cliniquement pertinentes pour prendre ses décisions. C'est une étape clé pour s'assurer que le comportement du modèle est cohérent, transparent, et potentiellement utile en pratique médicale.

5.3. Grad-CAM (Gradient-weighted Class Activation Mapping)

Pour mieux comprendre "où" notre modèle regarde lorsqu'il prend une décision, nous avons utilisé la méthode **Grad-CAM**. Cette technique génère une cartethermique superposée à l'image d'entrée, mettant en évidence les zones qui ont le plus influencé la prédiction. C'est un outil particulièrement utile pour visualiser les régions que le modèle considère comme importantes.

Voici les étapes que nous avons suivies :

- **Chargement du modèle**

Nous avons commencé par charger notre modèle de détection de cancer du sein, déjà entraîné, et l'avons mis en mode évaluation. En fonction des ressources disponibles, nous l'avons ensuite transféré sur le GPU ou resté sur le CPU.

- **Sélection de la couche cible**

Grad-CAM nécessite de cibler une couche convolutionnelle du réseau pour analyser l'activation des neurones. Nous avons choisi une couche située vers la fin de la partie convolutionnelle du modèle, car c'est à ce niveau que les caractéristiques les plus discriminantes sont généralement extraites.

- **Préparation de l'image**

Nous avons ensuite chargé une image d'exemple, l'avons convertie en format RGB et redimensionnée à 224×224 pixels. Cette image a été transformée en tenseur, puis mise sous forme de batch pour être compatible avec le modèle.

- **Génération de la heatmap Grad-CAM**

À l'aide de la bibliothèque pytorch-grad-cam, nous avons généré une carte d'activation en niveaux de gris. Cette heatmap reflète l'intensité avec laquelle chaque région de l'image a contribué à la décision du modèle.

- **Visualisation finale**

Nous avons superposé la heatmap sur l'image originale, ce qui nous a permis de voir clairement les zones "sur lesquelles le modèle s'est concentré". L'image annotée a ensuite été affichée aux côtés de l'image d'origine, facilitant ainsi la comparaison.

Grâce à cette méthode, nous avons pu juger si le modèle se focalisait sur des zones cohérentes d'un point de vue anatomique, ce qui est essentiel pour valider sa pertinence clinique et mieux comprendre son comportement.

6. *Les métriques de classification*

6.1.La matrice de confusion

Dans la matrice de confusion on croise les classes cibles réelles avec les classes prédites obtenues. Ceci nous donne le nombre d'instances correctement classées et mal classées.

		Négative	Positive
	Négative	**VN**	**FN**
	Positive	**FP**	**VP**

Tableau 3.3. Matrice de confusion pour la classification binaire.

- **VP** : vrais positifs sont le nombre d'instances positives correctement classifiées
- **FP** : faux positifs sont le nombre d'instances négatives et qui sont prédites comme positives.
- **FN** : faux négatifs sont le nombre d'instances positives classifiées comme négatives.
- **VN** : vrais négatifs sont le nombre d'instances négatives correctement classifiées.

À partir de la matrice de confusion on peut calculer plusieurs métriques qu'on va expliquer dans les sections suivantes (Benmaamar.O, 2022)

6.2. Les métriques

L'ensemble des métriques sont utilisées pour évaluer les méthodes d'apprentissage automatique. À partir de la matrice de confusion, de nombreuses mesures de performance du modèle peuvent être dérivées. (Benmaamar.O, 2022)

- **Accuracy**: correspond à tous les modèles correctement classés divisés par le nombre total de modèles. (Benmaamar.O, 2022)

$$\textbf{Accuracy} = \frac{\textbf{VP} + \textbf{VN}}{\textbf{VP} + \textbf{FP} + \textbf{FN} + \textbf{VN}}$$

- **Précision :** Cela définit l'exactitude du modèle en termes de prédiction . (Benmaamar.O, 2022)

$$\textbf{Précision} = \frac{\textbf{VP}}{\textbf{VP} + \textbf{FP}}$$

- **Recall** (Sensitivité): Cette mesure de performance implique comment différentes valeurs et variables indépendantes affectent une variable dépendante. (Benmaamar.O, 2022)

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

- **Le score F1 (F1 score) :** Peut être interprété comme une moyenne pondérée de la précision et la sensitivité, où un score F1 atteint sa meilleure valeur à 1 et son pire score à 0. Par conséquent, ce score prend en compte à la fois les cas faux positifs et les cas faux négatifs. Intuitivement, ce n'est pas aussi facile à comprendre que le taux de succès, mais F1 est généralement plus utile que le taux de succès, surtout si nous avons une distribution de classe inégale. Le taux de succès fonctionne mieux si les cas faux positifs et les cas faux négatifs ont une valeur similaire. Si la valeur des cas faux positifs et des cas faux négatifs est très différente, il est préférable d'examiner à la fois la précision et la sensitivité. Le score F1 est une métrique unique qui combine la sensitivité et la précisions en utilisant la moyenne harmonique. (Benmaamar.O, 2022)

7. Tableau comparatif des méthodes XAI utilisé :

Comment ça marche ?	Il attribue une importance à chaque pixel en se basant sur des concepts mathématiques solides (théorie des jeux).	Il regarde où le réseau « fait attention » en utilisant les gradients dans ses couches internes.	Il teste l'effet de cacher ou modifier des petites zones (superpixels) pour voir leur impact local.
Type d'explication	Très précise, pixel par pixel, locale.	Montre des zones importantes, plutôt qu'un détail pixel.	Explique par groupes de pixels (superpixels) proches.
À quoi ça ressemble ?	Une carte de chaleur avec des couleurs chaudes/froides qui montrent l'impact positif ou négatif.	Une surbrillance colorée sur l'image qui indique les régions clés.	Des zones découpées en morceaux colorés selon leur influence.

Facile à lire pour un médecin ?	Moyennement, parfois trop détaillé et difficile à interpréter rapidement.	Plutôt facile, car les zones surlignées correspondent bien à ce qu'on cherche.	Assez clair, les segments sont bien délimités et compréhensibles.
Fiabilité / robustesse	Théoriquement solide, mais peut être un peu bruité et moins stable parfois.	Plutôt robuste, mais dépend de la couche du réseau qu'on utilise.	Moyenne, dépend beaucoup de la qualité du découpage en superpixels.
Temps nécessaire	Relativement long, car il calcule beaucoup d'éléments.	Rapide, assez simple à générer.	Moyen, car il faut tester plusieurs modifications locales.
Niveau de détail	Très fin, jusqu'au pixel.	Moyen, montre plutôt des zones importantes.	Moyen, analyse par petits groupes de pixels.
Quand l'utiliser ?	Pour des analyses approfondies quand on veut vraiment comprendre chaque pixel.	Pour un diagnostic rapide et visuel, efficace en pratique.	Quand on a besoin d'expliquer des cas difficiles avec des détails locaux bien ciblés.

Tableau 3 .4. Comparatif des Méthodes d'Explication en Deep Learning Médical.

En nous appuyant sur le tableau de comparaison , on considère que chaque méthode a ses points forts . Cependant, Grad-CAM semble être la méthode la plus pratique et la plus intuitive pour une utilisation clinique, notamment parce qu'elle met rapidement en évidence les régions importantes de l'image. SHAP et LIME fournissent également des informations utiles, mais leur interprétation nécessite plus de prudence : SHAP est très précis mais complexe, tandis que LIME dépend fortement de la qualité de la segmentation de l'image. En fin de compte, le choix de la méthode dépend de vos besoins : rapidité, précision ou clarté.

8. Conclusion

Dans ce chapitre, nous avons décrit les étapes suivies pour réaliser notre étude, nous avons expliqué comment les données ont été préparées, quels outils ont été utilisés, et comment les

modèles ont été construits et interprétés. Chaque choix méthodologique a été fait avec le souci de répondre au mieux aux objectifs du projet. Cette approche nous guidera pour comprendre et discuter les résultats obtenus par la suite.

Chapitre

IV

Résultats expérimentaux

1. Introduction

Dans ce chapitre, nous avons marqué un tournant dans notre projet. Après avoir conçu, entraîné et testé différents modèles d'intelligence artificielle pour analyser des mammographies, il est temps de passer à une étape essentielle : comprendre ce que ces modèles valent vraiment, et comment ils prennent leurs décisions.

Nous allons donc, dans un premier temps, observer leurs performances. Ont-ils réussi à détecter correctement les cas de cancer ? Ont-ils fait beaucoup d'erreurs ? Pour y répondre, nous utiliserons des mesures couramment utilisées en médecine, comme la précision ou la sensibilité, afin d'évaluer leur fiabilité dans un contexte aussi sensible que le diagnostic du cancer du sein.

Mais cette évaluation ne s'arrête pas là. Car au-delà des chiffres, il est tout aussi important de savoir pourquoi un modèle donne telle ou telle réponse. Dans le domaine médical, cette transparence est indispensable : les professionnels doivent pouvoir comprendre et interpréter les décisions prises par l'algorithme. C'est là que les outils d'explicabilité entrent en jeu. Grâce à des méthodes comme Grad-CAM, SHAP ou LIME, nous pouvons visualiser les zones de l'image qui ont influencé la prédiction du modèle, et ainsi mieux saisir sa logique.

Ce chapitre est donc structuré autour de deux grands axes : d'abord l'analyse des performances techniques des modèles, puis l'exploration des explications qu'ils fournissent. Ces deux approches combinées nous permettront de mieux juger leur utilité réelle dans un cadre médical.

2. Evaluation des modèles de classification

Une fois nos modèles entraînés, nous avons évalué leurs performances pour vérifier leur fiabilité dans leurs prédictions. Pour cela, nous avons recours à la matrice de confusion, qui nous permet de visualiser précisément les bonnes classifications ainsi que les erreurs réalisées par chaque modèle.Voici les résultats de la matrice de confusion :

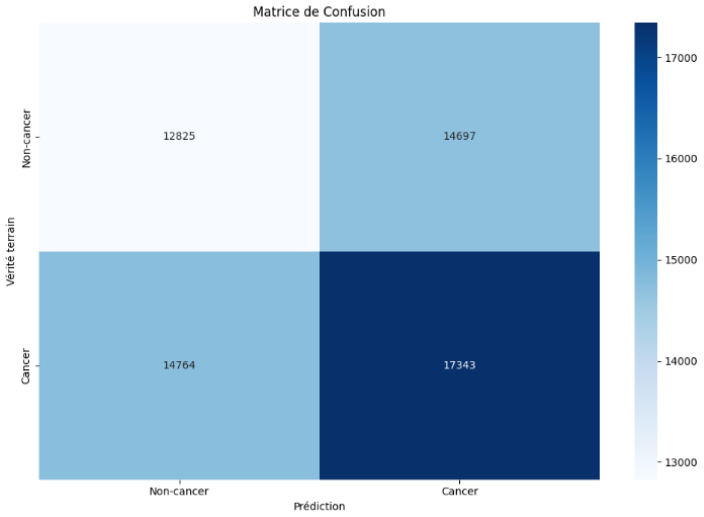

Figure 4.1. La matrice de confusion de modèle EfficientNetB0.

Vrai Positif (VP) :17343 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classecancer.

Faux Positif (FP) :14764 cas où le modèle a prédit à tort la classe cancer, alors que la classe réelle était non cancer.

Faux Négatif (FN) :14697 cas où le modèle a prédit la classe non cancer, alors que l'image appartenait en réalité à la classe cancer.

Vrai Négatif (VN) :12825 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classe non cancer.

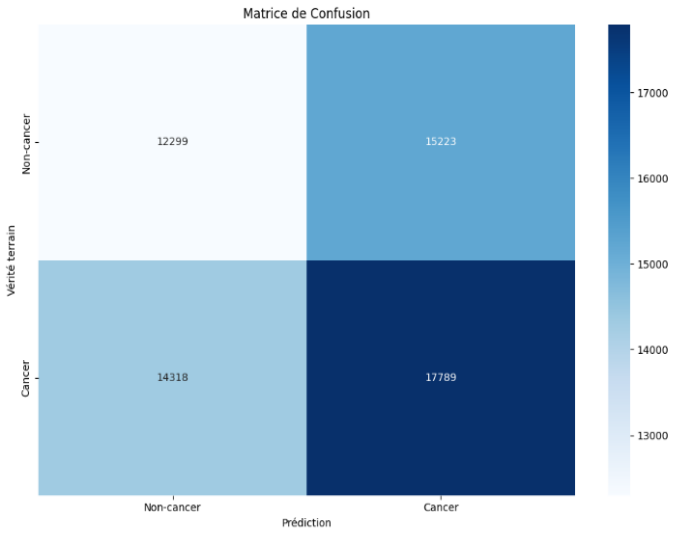

Figure 4.2. La matrice de confusion de modèle ResNet50.

Vrai Positif (VP) :17789 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classecancer.

Faux Positif (FP) :14318 cas où le modèle a prédit à tort la classe cancer, alors que la classe réelle était non cancer.

Faux Négatif (FN) :15223 cas où le modèle a prédit la classe non cancer, alors que l'image appartenait en réalité à la classe cancer.

Vrai Négatif (VN) :12299 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classe non cancer.

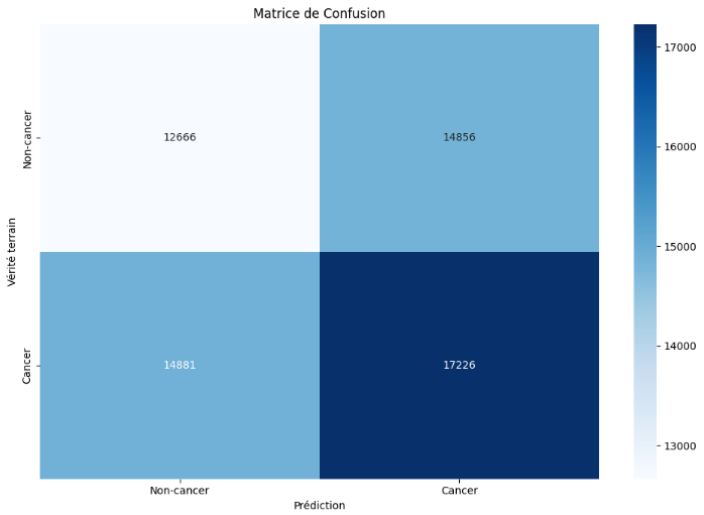

Figure 4.3. La matrice de confusion de modèle DensNet121.

Vrai Positif (VP) :17226 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classecancer.

Faux Positif (FP) :14881 cas où le modèle a prédit à tort la classe cancer, alors que la classe réelle était non cancer.

Faux Négatif (FN) :14856 cas où le modèle a prédit la classe non cancer, alors que l'image appartenait en réalité à la classe cancer.

Vrai Négatif (VN) :12666 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classe non cancer.

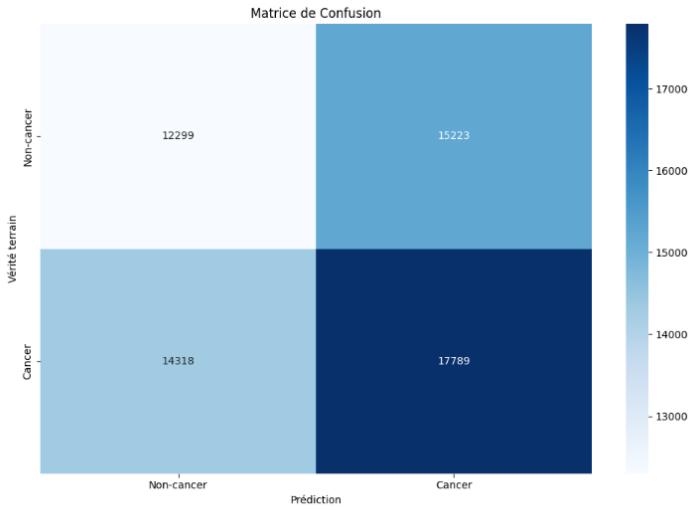

Figure 4.4. La matrice de confusion de modèle ConvNeXt-Tiny .

Vrai Positif (VP) :17789 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classecancer.

Faux Positif (FP) :14318 cas où le modèle a prédit à tort la classe cancer, alors que la classe réelle était non cancer.

Faux Négatif (FN) :15223 cas où le modèle a prédit la classe non cancer, alors que l'image appartenait en réalité à la classe cancer.

Vrai Négatif (VN) :12299 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classe non cancer.

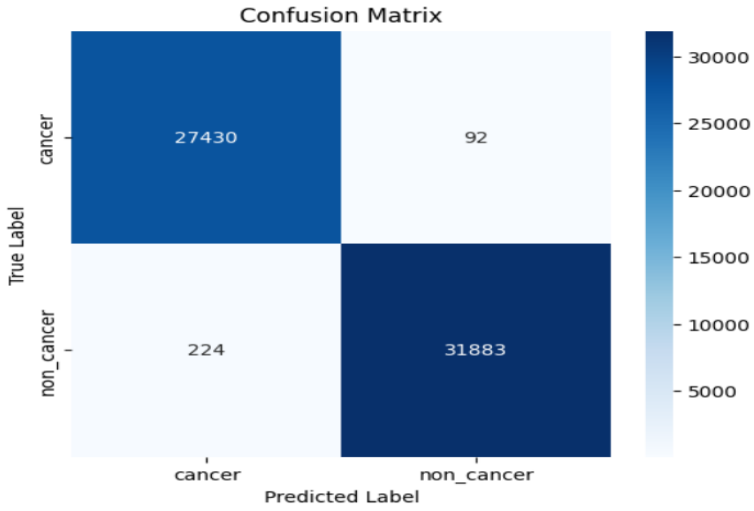

Figure 4.5. La matrice de confusion de modèle CNN.

Vrai Positif (VP) :31883 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classecancer.

Faux Positif (FP) :244 cas où le modèle a prédit à tort la classe cancer, alors que la classe réelle était non cancer.

Faux Négatif (FN) :92 cas où le modèle a prédit la classe non cancer, alors que l'image appartenait en réalité à la classe cancer.

Vrai Négatif (VN) :27430 cas dans lesquels le modèle a correctement identifié une image comme appartenant à la classe non cancer.

Le tableau suivant en résume les résultats, en indiquant le nombre de vrais positifs (VP), faux positifs (FP), faux négatifs (FN) et vrais négatifs (VN).

EfficientNetB0	17343	14764	14697	12825
ResNet50	17789	14318	15223	12299
DenseNet121	17226	14881	14856	12666
ConvNeXt-Tiny	17789	14318	15223	12299
CNN	31883	224	92	27430

Tableau 4.1. Tableau résumé les matrices de confusion par modèle.

D'après les matrices de confusion, nous observons que les modèles préentraînés, tels qu'EfficientNetB0, ResNet50, DenseNet121 et ConvNeXt-Tiny, font beaucoup d'erreurs, avec un grand nombre de faux positifs et de faux négatifs. Cela montre qu'ils ont encore du mal à bien faire la distinction entre les classes. En revanche, le modèle CNN personnalisé excelle clairement, avec très peu d'erreurs : 224 faux positifs et seulement 92 faux négatifs. Il produit également le plus grand nombre de prédictions correctes, démontrant ainsi son excellente adaptation à notre ensemble de données

Les performances des modèles ont également été évaluées à l'aide de trois métriques clés : la précision, le rappel et le F1-score, présentées ci-dessous pour chaque modèle.

EfficienNetB0	**99.97**	**54.13**	**54.02**	**54.07**
DenseNet121	**99.94**	**53.69**	**53.65**	**53.67**
ResNet50	**99.36**	**53.99**	**55.41**	**54.64**
ConvNexT-Tiny	**99**	**53.89**	**55.41**	**54.64**
CNN	**99.47**	**99.71**	**99.30**	**99.51**

Tableau 4.2. Un tableau résume les résultats obtenus par les cinq modèles teste .

À première vue, les cinq modèles testés semblent exceptionnellement performants, avec une précision supérieure à 99 %. Cependant, en termes de précision, de rappel et de scores F1, nous observons des différences significatives entre les modèles testés. Par exemple, le CNN hautement personnalisé a atteint des niveaux de performance extrêmement élevés, avec une précision de 99,3 %, un rappel de 99,7 % et un score F1 de 99,5 %. Ces résultats reflètent une excellente capacité à détecter les cas positifs tout en minimisant les erreurs de classification, ce que nous recherchions.

Pour d'autres modèles, ces résultats sont plus mitigés. Par exemple, ResNet50 atteint une précision de 55,4 % et un rappel de 53,9 %, ce qui démontre la difficulté de détecter avec précision tous les cas positifs. DenseNet121 se classe dans une position similaire, avec des scores autour de 53,7 %. Ces résultats indiquent que ces modèles peinent à distinguer les images infectées des images non infectées, ce qui limite leur fiabilité dans le contexte médical.

3. Interprétation du modèle avec les méthodes de XAI

Bien que notre modèle CNN personnalisé ait obtenu d'excellents résultats, il reste difficile de comprendre précisément comment il prend ses décisions. Nous avons donc utilisé des techniques d'explicabilité (XAI) telles que LIME, SHAP et Grad-CAM pour mieux comprendre son fonctionnement. Ces méthodes identifient les régions ou les éléments d'image

qui influencent le plus les prédictions du modèle, vérifiant ainsi la cohérence et la clarté de son raisonnement.Chaque méthode testée s'est révélée apporter un éclairage différent et complémentaire sur les prédictions du modèle.

3.1. Shap (Shapley Additional Interpretations)

Tout d'abord, nous avons utilisé la méthode SHAP (Shapley Additional Interpretations) pour comprendre comment notre modèle CNN prend ses décisions à partir des images mammographiques. Les figures ci-dessus présentent deux exemples tirés de notre base de données. Pour chaque cas, l'image de gauche montre la mammographie originale, tandis que l'image de droite présente une carte générée par SHAP, mettant en évidence les régions ayant le plus influencé les prédictions du modèle. Les zones rouges représentent les régions ayant conduit le modèle à prédire un cas de cancer, tandis que les zones bleues indiquent les régions ayant influencé un diagnostic « non cancéreux ».

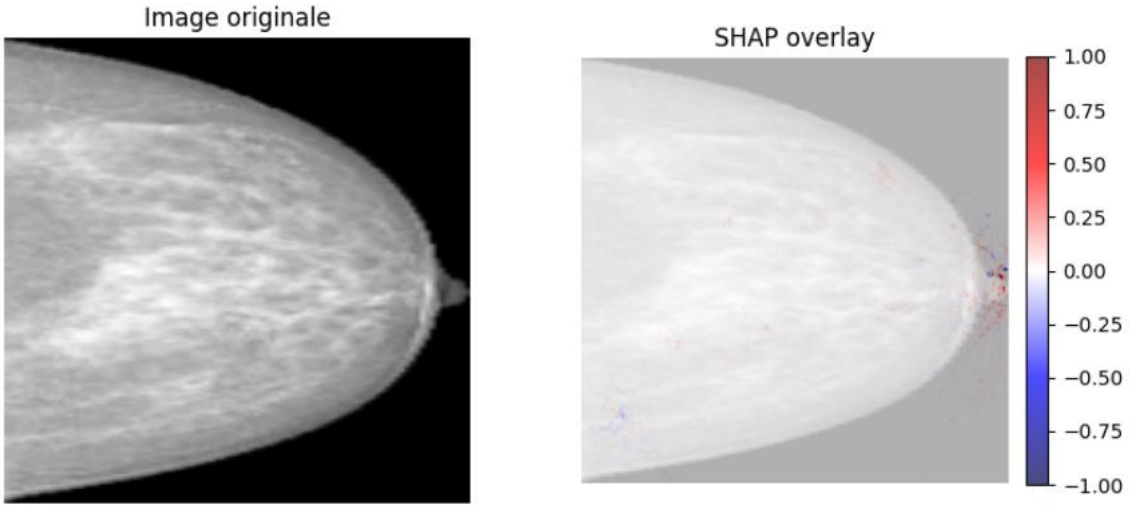

Figure 4.6.un exemple d'interprétation SHAP appliquée à une image mammographique.

Dans le premier exemple, la carte SHAP fait apparaître plusieurs zones rouges bien marquées, en particulier autour du mamelon. Ces régions indiquent que le modèle a identifié des caractéristiques visuelles qu'il associe fortement à la présence d'un cancer. En se basant sur ces activations, le modèle a donc classé cette image dans la catégorie « cancer ».

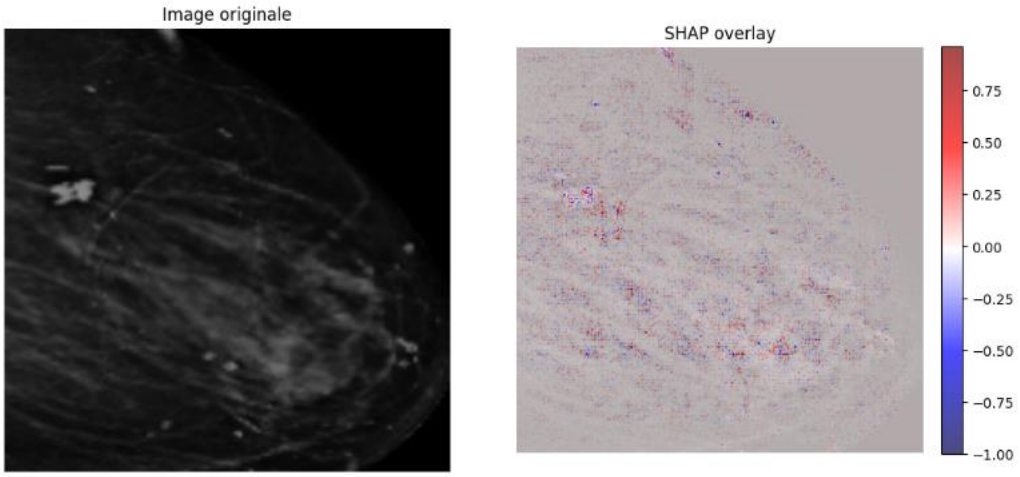

Figure 4.7. Deuxième exemple d'interprétation SHAP appliquée à une image mammographique.

Dans le deuxième exemple, on observe également de nombreuses zones rouges, surtout vers le centre de l'image, ce qui renforce encore une fois l'idée d'une prédiction orientée vers la classe « cancer ». Ce type de visualisation montre tout l'intérêt de la méthode SHAP : elle ne se contente pas de donner une réponse, mais met en lumière les éléments précis de l'image qui ont influencé la décision du modèle. Cela rend le processus beaucoup plus transparent et facilite l'interprétation des résultats, en rapprochant le fonctionnement du modèle du raisonnement habituel des professionnels de santé.

3.2. Grad-Cam (Gradient-weighted Class Activation Mapping)

Nous avons également utilisé Grad-CAM (Gradient-weighted Class Activation Mapping) pour générer une carte thermique superposée à l'image originale, mettant en évidence les régions les plus actives dans les couches profondes du réseau neuronal. Les figures ci-dessous nous permettent de mieux comprendre les éléments de l'image que le modèle considère comme importants pour ses prédictions.Sur ces visualisations, les teintes rouges et jaunes indiquent les zones que le modèle a jugées les plus significatives pour établir son diagnostic. À l'inverse, les zones bleues correspondent à des régions qu'il a considérées comme peu pertinentes, voire sans influence sur sa décision.

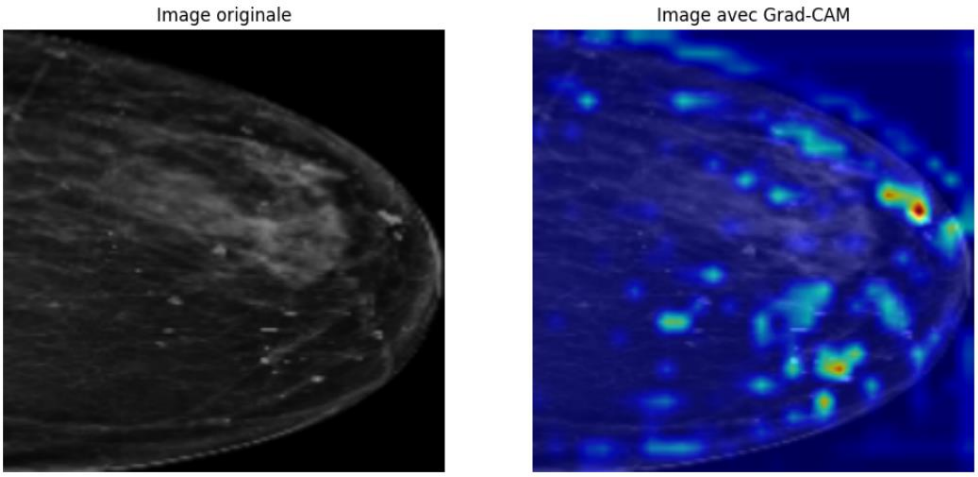

Figure 4.8. Un exemple d'interprétation Grad-CAM appliquée à une image mammographique. Dans le premier exemple, on observe plusieurs zones d'activation bien marquées, réparties de façon diffuse, notamment vers les bords de la mammographie. Cette distribution étalée des activations suggère que le modèle ne s'est pas concentré sur une seule région spécifique, mais qu'il a plutôt pris en compte différents éléments visuels dispersés à travers l'image. Ce type de réponse peut traduire la présence de signes pathologiques plus discrets ou étendus, comme des microcalcifications disséminées, des distorsions architecturales ou des zones d'hyperdensité suspectes.

Sur le plan algorithmique, cela signifie que le modèle a exploité plusieurs indices répartis pour formuler sa prédiction, ce qui montre une forme d'analyse plus globale de l'image. Un tel comportement est typique des cas plus complexes, où l'anomalie ne se résume pas à une masse nette ou centrée, mais à une combinaison de signaux subtils. C'est précisément dans ces situations que les méthodes d'explicabilité, comme Grad-CAM, prennent tout leur sens : elles offrent une lecture visuelle du raisonnement du modèle, permettant ainsi de mieux comprendre comment il parvient à ses conclusions.

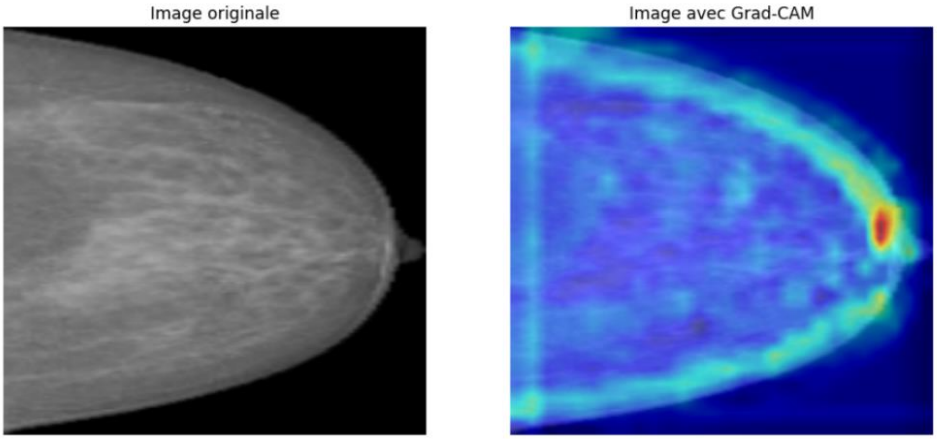

Figure 4.9. Deuxième exemple d'interprétation Grad-CAM appliquée à une image mammographique.

Dans la deuxième paire d'images, l'activation est beaucoup plus concentrée. Le modèle semble vraiment se focaliser sur la zone autour du mamelon, où la carte Grad-CAM montre une couleur rouge intense. Cela signifie qu'il a repéré dans cette partie de l'image un élément important. Cette concentration montre que le modèle s'appuie sur un indice clair pour prendre sa décision, ce qui rend sa prédiction plus précise. Quand l'activation est aussi nette et ciblée, cela indique souvent que le modèle a trouvé ce qui compte vraiment, sans se disperser ailleurs. C'est aussi pour ça que les cartes Grad-CAM sont utiles : elles nous aident à voir où le modèle "regarde" pour arriver à son résultat.

3.3. LIME (Local Interpretable Model-Agnostic Explanations)

Enfin, nous avons utilisé la méthode LIME(Local Interpretable Model-Agnostic Explanations) qui offre une explication locale en identifiant les parties de l'image(superpixels) qui ont influencé la prédiction du modèle. Les superpixels en vert montrent les zones qui ont soutenu la prédiction de la classe visée par exemple : « non-cancer », tandis que ceux en rouge indiquent les régions qui ont favorisé la prédiction de l'autre classe « cancer ». Grâce à cette représentation visuelle, il devient plus simple de comprendre quels éléments de l'image ont joué un rôle clé dans la décision du modèle.

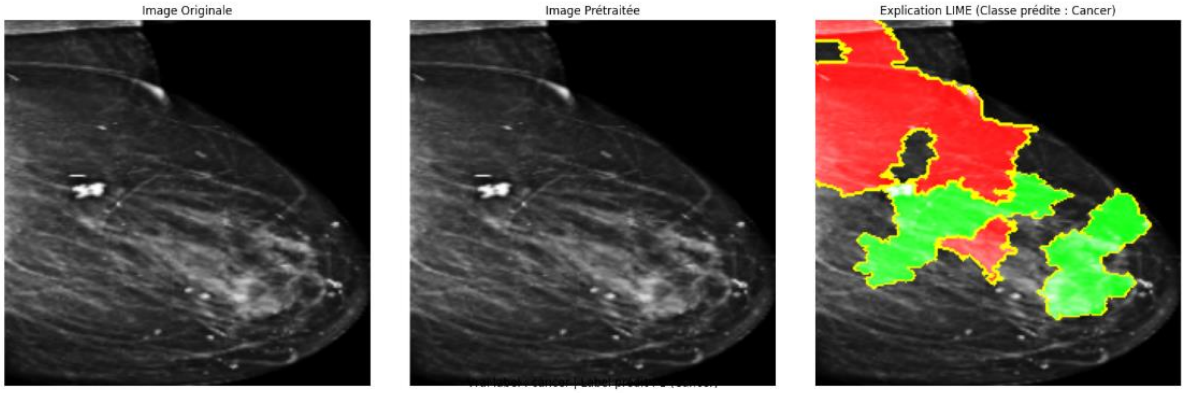

Figure 4.10. Un exemple d'interprétation LIME appliquée à une image mammographique.

Dans le premier exemple,LIME a été utilisé sur un cas que le modèle a prédit comme positif au cancer. L'image explicative révèle plusieurs superpixels rouges, surtout dans la partie haute de la glande mammaire. Cela signifie que le modèle a identifié dans ces zones des signes visuels typiques du cancer, comme des motifs irréguliers ou des anomalies de texture. On remarque aussi quelques zones en vert, indiquant que certaines parties de l'image ont plutôt soutenu une prédiction « non-cancer », mais pas assez pour changer la décision finale. Ce contraste montre bien l'importance relative des différentes régions que le modèle a prises en compte.

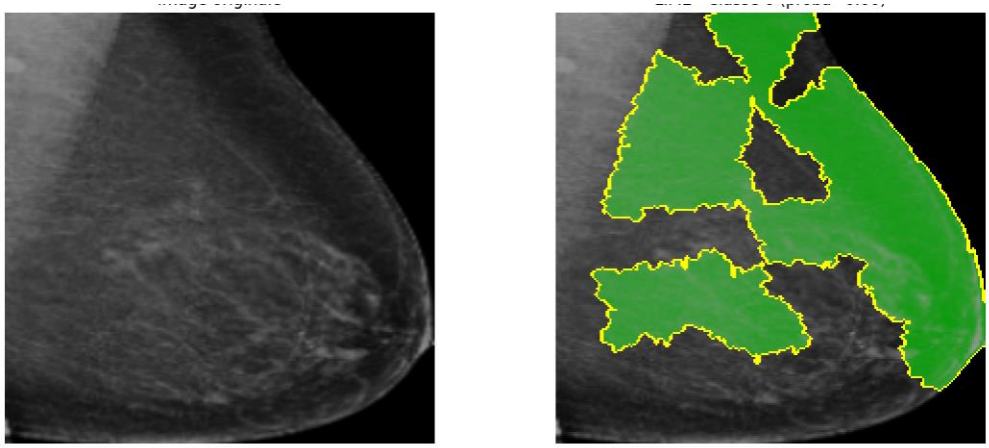

Figure 4.11. Deuxième exemple d'interprétation LIME appliquée à une image mammographique.

Dans un deuxième exemple, la prédiction du modèle est « non-cancer », et la visualisation LIME vient confirmer ce résultat. On y voit surtout des superpixels verts répartis un peu partout sur l'image, sans aucune zone rouge. Cela signifie que, selon le modèle, l'image ne montre pas de signes suspects typiques d'une lésion cancéreuse. Cette correspondance entre la prédiction et l'explication visuelle renforce la confiance que l'on peut avoir dans le modèle.

Les trois méthodes d'explicabilité (SHAP, Grad-CAM et LIME) nous aident à mieux comprendre comment le modèle prend ses décisions, mais chacune le fait à sa façon. **SHAP** donne une analyse très détaillée en mettant en lumière les zones de l'image qui ont joué pour ou contre la prédiction finale. **Grad-CAM**, de son côté, nous montre où le modèle a concentré son attention grâce à une carte de chaleur superposée à l'image, ce qui permet de visualiser ses « zones d'intérêt ». Enfin, **LIME** découpe l'image en morceaux appelés superpixels, et colore ceux qui ont influencé la décision en vert ou en rouge, selon leur impact.

4. Conclusion

Nous concluons que les méthodes d'interprétation jouent un rôle essentiel pour faciliter la compréhension des décisions prises à l'aide de modèles. Chacune offre une perspective différente : Grad-CAM se caractérise par sa simplicité et sa clarté visuelle, ce qui en fait un outil performant pour une utilisation clinique. En revanche, SHAP et LIME permettent des analyses plus détaillées, même si elles nécessitent une interprétation plus poussée. En combinant ces méthodes, nous pouvons fournir des interprétations plus fiables, plus claires et plus utiles aux professionnels de santé.

Conclusion générale

L'intelligence artificielle prend une place de plus en plus importante dans le domaine médical, et son utilisation pour aider au dépistage du cancer du sein en est un exemple particulièrement prometteur. À travers ce projet, nous avons voulu montrer qu'il est possible de concevoir des modèles de classification automatique capables d'analyser des mammographies avec un très haut niveau de précision, tout en restant compréhensibles et interprétables par les professionnels de santé.

Nous avons d'abord pris le temps de comprendre en profondeur les bases de l'IA et de l'explicabilité, des domaines à la fois techniques et passionnants, mais souvent complexes. Cette première étape était essentielle pour aborder ensuite, avec sérieux et sensibilité, le contexte médical du cancer du sein : un sujet qui touche profondément, et qui rappelle que derrière chaque image se cache une personne, une histoire, une vie.

Sur le plan technique, plusieurs architectures de réseaux de neurones ont été testées, comparées et évaluées à l'aide du dataset ***''Final-Final-RSNA-Breast-Cancer-Dataset ''***. Nous avons également développé un modèle personnalisé, qui s'est avéré être le plus performant, avec une précision de 99,47 %. Ces résultats montrent clairement que les technologies actuelles sont capables de grandes choses — à condition d'être bien utilisées.

Mais au-delà des performances, ce qui nous a semblé essentiel, c'est de rendre ces modèles explicables. En effet, dans un contexte médical, une prédiction seule ne suffit pas : il faut pouvoir justifier cette prédiction, la visualiser, la comprendre. C'est dans cet esprit que nous avons exploré plusieurs méthodes d'explicabilité visuelle, comme Grad-CAM, SHAP et LIME. Grad-CAM s'est révélée être la plus intuitive, car elle permet de voir clairement quelles zones de l'image ont guidé la décision du modèle, ce qui représente une vraie aide pour les radiologues.

En fin de compte, ce projet a été bien plus qu'un simple exercice technique. Il a été l'occasion de réfléchir à l'impact réel que peuvent avoir les algorithmes sur la vie des gens. Cependant, il reste encore beaucoup à faire : intégrer les données cliniques, développer des approches plus interprétables et tester ces outils dans des contextes cliniques réalistes.

Références

Alejandro Barredo Arrieta, N., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2019). *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI* (Version 2, 26 décembre 2019). arXiv. https://arxiv.org/pdf/1910.10045

Barka, Z. (2022). *L'intelligence artificielle entre réalité et futur*. Université Kasdi Merbah Ouargla. https://dspace.univ-ouargla.dz/jspui/bitstream/123456789/37016/1/ZITOUNI%20-%20BARKA%20.pdf

Barka, Z. (2023). *Analyse du système de reconnaissance automatique de la parole*. Université Kasdi Merbah Ouargla. Consulté le 28 avril 2025, à partir de https://dspace.univ-ouargla.dz/jspui/bitstream/123456789/37016/1/ZITOUNI%20-%20BARKA%20.pdf

Bellel Maroua, S., & Boukhenaf, Y. (2018, 1 juillet). *Étude statistique, histologique et moléculaire du cancer du sein dans la région de Constantine* [Mémoire de fin d'études, Université des Frères Mentouri Constantine]. https://fac.umc.edu.dz/snv/bibliotheque/biblio/mmf/2018/Etude%20statistique,%20histologique%20et%20mol%C3%A9culaire%20du%20cancer%20du%20sein%20dans%20la%20r%C3%A9gion%20de%20Constantine..pdf

Benmaamar, O. (2022). *Breast cancer classification using transfer learning* [Mémoire de fin d'études, Université 8 Mai 1945 - Guelma]. DSpace Université de Guelma. https://dspace.univ-guelma.dz/jspui/bitstream/123456789/14927/1/BENMAAMAR_OUSSAMA_F5.pdf

Benmaamar, O. (2023, juin). *Cancer du sein : étude et analyse des facteurs de risque* [Mémoire de fin d'études, Université 8 Mai 1945 Guelma]. DSpace Université 8 Mai 1945 Guelma. https://dspace.univ-guelma.dz/jspui/bitstream/123456789/14927/1/BENMAAMAR_OUSSAMA_F5.pdf

Zara, I. (2019).L'intelligence artificielle principe, outils et objectifs. [Mémoire, Université d'Annaba]. Biblio Ingéniorat. https://biblio.univ-annaba.dz/ingeniorat/wp-content/uploads/2019/09/Zara-Islem.pdf

Bounia, L. (2023, 22 décembre). *Modèles formels pour l'IA explicable : des explications pour les arbres de décision* [Mémoire de master, Université d'Artois, France]. Centre de Recherche en Informatique de Lens (CRIL – CNRS UMR 8188).

Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). *A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME*. *Advanced Intelligent Systems, 7*, Article 2400304. https://doi.org/10.1002/aisy.202400304

Centre International de Recherche sur le Cancer (CIRC). (2022). *Cancer Today – Visualisation en camembert* [Visualisation de données]. Consulté le 18 février 2025, sur

https://gco.iarc.who.int/today/en/dataviz/pie?mode=cancer&types=1&sexes=2&populations=900

EITCA Academy. (s.d.). *Est-il possible d'utiliser Kaggle pour télécharger des données financières et effectuer une analyse statistique et des prévisions ?* Récupéré sur https://fr.eitca.org/artificial-intelligence/eitc-ai-gcml-google-cloud-machine-learning/advancing-in-machine-learning/data-science-project-with-kaggle/is-it-possible-to-use-kaggle-to-upload-financial-data-and-perform-statistical-analysis-and-forecasting-u

Juniper Networks. (s.d.). *IA explicable (XAI)*. Consulté le 1er juillet 2025, sur Juniper Networks :https://www.juniper.net/fr/fr/research-topics/what-is-explainable-ai-xai.html

Fondation pour la Recherche Médicale (FRM). (2023). *Focus sur le cancer du sein*. https://www.frm.org/fr/maladies/recherches-cancers/cancer-du-sein/focus-cancer-sein

Hallali, H. (2022, 6 juillet). *Estimation de la pose de la caméra basée sur un réseau neuronal convolutif* [Mémoire d'ingénieur, École Nationale Polytechnique d'Alger].

Hicks, S. A., Langøien, L. J., Riegler, M. A., Halvorsen, P., & Eskeland, S. L. (2020, 24 septembre). *Explaining Deep Neural Networks for Knowledge Discovery in Electrocardiogram Analysis*. arXiv. https://arxiv.org/pdf/2009.11698

Lasnier, A. (2024). *Le cancer du sein : physiopathologie, prédispositions génétiques et stratégies thérapeutiques* [Mémoire de master, DUMAS]. DUMAS - Dépôt Universitaire de Mémoires Après Soutenance. https://dumas.ccsd.cnrs.fr/dumas-04425205v1/file/LASNIER%20Ana%C3%AFs.pdf

Lazhar, M. (2022). *L'impact de l'intelligence artificielle sur le management des ressources humaines* [Mémoire de master, Université de Guelma]. https://dspace.univ-guelma.dz/jspui/bitstream/123456789/13224/1/Memoire%20MAKHLOUF_Lazhar_F.pdf__%20%281%29.pdf

MathWorks. (s.d.). *efficientnetb0 (Deep Learning Toolbox)*. Récupéré sur https://fr.mathworks.com/help/deeplearning/ref/efficientnetb0.html

MathWorks. (s.d.). *resnet50 (Deep Learning Toolbox)*. Récupéré sur https://www.mathworks.com/help/deeplearning/ref/resnet50.html

Mezouaghi, A., & Niati, R. (2020). *Classification des images de mammographie*. http://dspace.univ-tlemcen.dz/bitstream/112/17306/1/Ms.GBM.Niati%2BMezouaaghi.pdf

Mooney, C., Mazo, C., Al-Azab, A., Dineen, Y., Wang, L., & Gallagher, W. M. (2011). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences, 11*(11), 5088. https://doi.org/10.3390/app11115088

Neffaf, R., Keniouche, N., & Bougherara, O. (2022, 25 juin). *La prédisposition génétique du gène BRCA1 au cancer du sein* [Mémoire de fin d'études, Faculté des Sciences de la Nature et de la Vie, Université Mentouri]. https://fac.umc.edu.dz/snv/bibliotheque/biblio/mmf/2022/La%20pr%C3%A9dispositio

n%20G%C3%A9n%C3%A9tique%20du%20g%C3%A8ne%20BRCA1%20au%20Cancer%20du%20sein.pdf

Nihad, B., & Hadjer, L. (2019). *Identification faciale en 3D pour le contrôle d'accès aux zones sensibles* [Mémoire de master, Université Abdelhamid Ibn Badis Mostaganem].

Sari, Y. (2023). *Détection et classification des anomalies dans les mammographies en utilisant les techniques de Deep Learning* [Mémoire de master, Université Badji Mokhtar Annaba].

Sharma, S. (2023, 28 novembre). *EfficientNet architecture*. GeeksforGeeks. https://chatgpt.com/c/6810ddc3-2ee0-8000-b531-d58409e20056

van Rossum, G., & Python Development Team. (n.d.). *Python is an experiment in how much freedom programmers need. Too much freedom and nobody can read another's code; too little and expressiveness is endangered*. Python.org. https://www.python.org/doc/essays/blurb/