République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE
Présenté en vue de l'obtention du diplôme
Master en informatique
Spécialité : ingénierie d'informatique Décisionnel

THEME

Proposition d'une technique efficace pour l'analyse du big data basée sur l'intelligence artificielle.

Présenté par :

DEROUAZ Nour El houda

KEDJOUTI Ahlem

Soutenu publiquement le : 22/06/2023

Devant le jury composé de :

Président : Dr. Attia safa

Examinateur : Dr. Mohdeb Djamila

Encadreur : Dr. SENOUCI Oussama

2022/2023

“Je n’ai jamais rêvé de succès,
J’ai travaillé pour ça.”

- Estee Lauder -

Remerciement

Nous exprimons notre reconnaissance envers notre grand Être suprême d'avoir béni notre équipe de santé, de ressources, de volonté, de courage et d'opportunités nécessaires pour mener à bien cette étude et l'achever avec succès.

En premier lieu, nous tenons à souligner que ce travail n'aurait pas atteint un tel niveau de qualité sans l'aide précieuse et l'encadrement du **Dr. Oussama SENOUCI**. Nous tenons à vous exprimer notre gratitude spéciale pour votre encadrement dévoué tout au long de notre projet. Votre expertise et vos conseils éclairés ont été inestimables. Votre disponibilité constante, votre patience et votre volonté de partager vos connaissances ont grandement contribué à notre réussite. Nous vous sommes sincèrement reconnaissants pour votre implication et votre soutien inconditionnel.

Nous souhaitons également remercier notre amie, le **Dr. Aasma Trakia**, pour leur précieuse assistance dans la compréhension de certains aspects du domaine médical.

Nos remerciements s'adressent également à tous nos professeurs qui ont joué un rôle essentiel dans notre parcours académique. Votre enseignement passionné, vos encouragements et vos conseils avisés nous ont permis de progresser dans notre discipline et de développer nos compétences.

Enfin, nous exprimons notre profonde gratitude envers notre famille qui nous a toujours soutenus et a contribué à notre formation à tous les niveaux d'études.

Dédicace

Ce travail est dédie à

***mes chers parents**, mes piliers, mes premiers soutiens et ma plus grande force. Je vous remercie infiniment pour votre présence, votre soutien, votre aide financière, et surtout votre amour. Vous n'avez jamais douté de moi, et j'espère sincèrement que vous êtes fiers de moi aujourd'hui.car tout ce que j'ai accompli, c'est grâce à vous.*

*À mon frère **Abdel rahmane** et ma petite sœur **Imane** ,*

*À chaque membre de **ma Famille** , surtout **mes grands-parents** ,*

*À **mes amis**, qui ont été mes compagnons de route durant ces années d'études,*

Merci d'avoir partagé ces moments inoubliables avec moi,

*À tous **mes chers professeurs**.*

⋆ ***Nour El Houda*** ⋆

Dédicace

Ce mémoire est dédié à exprimer ma gratitude envers Dieu pour toute l'énergie qu'Il m'a accordée durant ces cinq années, hamdoulillah. J'aimerais dédier mon travail à toutes les personnes qui m'ont apporté leur aide tout au long de mon parcours universitaire. En particulier, je tiens à remercier mes chers parents, mes frères, ma sœur, mon fiancé ainsi que toute ma famille, sans qui je n'aurais jamais pu réaliser ce travail.

⋆ ***Ahlem*** ⋆

Résumé

Ce travail propose une technique novatrice d'analyse du Big Data basée sur le deep learning pour améliorer la détection du cancer du sein. Le Big Data représente un ensemble massif de données générées quotidiennement à partir de diverses sources, et la détection précoce du cancer du sein est cruciale pour améliorer les chances de guérison.

L'utilisation du deep learning, en particulier des réseaux de neurones convolutionnels (CNN), offre de grandes opportunités pour extraire des informations significatives à partir d'images médicales telles que les mammographies. Cette approche permet d'automatiser le processus de détection et de classification des anomalies, ce qui peut contribuer à réduire les erreurs humaines et à accélérer le diagnostic.

Ce mémoire explore les différentes étapes du traitement du Big Data pour développer une méthode générale applicable à l'analyse de grandes quantités de données médicales. Nous mettons en évidence l'utilisation des CNN pour extraire des caractéristiques pertinentes à partir des images de mammographie et pour effectuer une classification précise des tissus mammaires.

Une étude de cas est présentée pour évaluer l'efficacité de la méthode proposée. Les résultats obtenus démontrent une amélioration significative de la détection du cancer du sein par rapport aux méthodes traditionnelles. Cette approche ouvre des perspectives prometteuses pour le dépistage précoce et précis du cancer du sein, ce qui peut avoir un impact positif sur le pronostic et la survie des patients.

Mots clés : Big Data, Deep Learning, Réseaux de neurones convolutifs (CNN), Cancer du sein, Détection précoce, Mammographie.

Abstract

This dissertation proposes an innovative technique for Big Data analysis based on deep learning to enhance breast cancer detection. Big Data represents a massive set of data generated daily from various sources, and early detection of breast cancer is crucial for improving the chances of recovery.

The utilization of deep learning, particularly convolutional neural networks (CNN), presents significant opportunities to extract meaningful information from medical images such as mammograms. This approach automates the process of anomaly detection and classification, thereby reducing human errors and expediting diagnosis.

The thesis explores the different stages of Big Data processing to develop a general method applicable to the analysis of large quantities of medical data. We highlight the use of CNNs to extract relevant features from mammography images and perform accurate classification of breast tissues.

A case study is presented to evaluate the effectiveness of the proposed method. The results demonstrate a substantial improvement in breast cancer detection compared to traditional methods. This approach holds promising prospects for early and precise breast cancer screening, which can have a positive impact on patients' prognosis and survival.

Keywords : Big Data, Deep Learning, Convolutional Neural Networks (CNN), Breast Cancer, Early Detection, Mammography.

ملخص

تقدم هذه الرسالة تقنية مبتكرة لتحليل البيانات الضخمة باستخدام التعلم العميق لتعزيز اكتشاف سرطان الثدي. تمثل البيانات الضخمة مجموعة ضخمة من البيانات التي يتم إنشاؤها يومياًا من مصادر مختلفة ، ويعتبر الكشف المبكر عن سرطان الثدي أمراً حاسماًا لتحسين فرص الشفاء.

يوفر استخدام التعلم العميق ، وبالأخص الشبكات العصبية التكاثفية ، فرصاًا كبيرة لاستخلاص معلومات ذات مغزى من الصور الطبية مثل صور الأشعة السينية للثدي. تسهم هذه الطريقة في تسريع عملية اكتشاف الاختلالات والتصنيف ، مما يقلل من الأخطاء البشرية ويسرع عملية التشخيص.

تستكشف الرسالة المراحل المختلفة لمعالجة البيانات الضخمة لتطوير طريقة عامة يمكن تطبيقها على تحليل كميات كبيرة من البيانات الطبية. نسلط الضوء على استخدام الشبكات العصبية التكاثفية لاستخلاص سمات ذات صلة من صور الأشعة السينية للثدي وإجراء تصنيف دقيق لأنسجة الثدي.

تم تقديم دراسة حالة لتقييم فعالية الطريقة المقترحة. تظهر النتائج تحسيناًا كبيراً في كشف سرطان الثدي مقارنة بالأساليب التقليدية. تتيح هذه الطريقة آفاقاًا واعدة لفحص سرطان الثدي المبكر والدقيق ، مما يمكن أن يكون له تأثير إيجابي على تشخيص ونجاح علاج المرضى.

الكلمات المفتاحية: البيانات الضخمة ، التعلم العميق ، الشبكات العصبية التكاثفية سرطان الثدي ، الكشف المبكر ، الأشعة السينية للثدي.

Table des matières

Liste des abréviations

AI Artificial Intelligence

SGBD Systèmes de Gestion de Bases de Bonnées

DAD domaine de l'analyse de données

ML Machine learning

GPS Global Positioning System

XML extensible markyp language

CSV comma separated values

DL Deep learning

5V Volume Vélocité Variété Véracité valeur

SSL Semi-supervised Learning

SVM Support Vector Machines

DT Decision tree

CNN Convolutional Neural Networks

RNN Convolutional Neural Networks

LSTM Long short-term memory

GRU Gated Recurrent Units

SAKs Autoencoder et Stacked Autoencoders

DBN Deep Belief Network

GPU Graphics Processing Unit

BDA-MC Big data analytic diabetics using map reduce and classification techniques

ML-SBD Applying spark based machine learning model on streaming big data for health status prediction

RAC-BD Earthquake Prediction in California Using Regression Algorithms and Cloud-based Big Data Infrastructure

EBD-C Fast and effective Big Data exploration by clustering

BDML-CP Big data and machine learning for crop protection.

CLUBS-P Clustering of Large and Unbalanced Biological Data Sets based on Pseudo-relevance feedback

MEFASD-BD Multi-objective Evolutionary Fuzzy Algorithm for Subgroup Discovery in Big Data environments - A MapReduce solution

SSL-G Semi-Supervised Learning with Graphs

MCA-MO Some methods for classification and analysis of multivariate observations

NN-ACF Neural Networks—A Comprehensive Foundation

in multimedia big data for mobile internet

DL-BDA Deep learning in big data analytics : A comparative study

SDL-BDA A survey on deep learning in big data analytics

PDRN-BDA processing framework based on improved distributed recurrent neural network variants with fasttext for social big data analytics

DL-BCDSM Deep Learning to Improve Breast Cancer Detection on Screening Mammography **CAD** Computer-Assisted Detection and Diagnosis

ROI Region Of Interest

CBIS-DDSM Curated Breast Imaging Subset of the Digital Database for Screening Mammography

AUC Area Under the Curve

FFDM Full-feld Digital Mammography

ACR American College of Radiology

Table des figures

Liste des tableaux

List of Algorithms

Chapitre 1

Introduction Générale

Dans le cadre de notre projet de fin d'études en Master 2 Ingénierie d'informatique Décisionnelle, nous avons entrepris une étude sur un sujet essentiel : **la proposition d'une technique basée sur l'intelligence artificielle pour une analyse efficace du big data**. Cette introduction mettra en avant le contexte et la problématique de notre mémoire, ainsi que les objectifs et les contributions de notre travail. Enfin, nous présenterons l'organisation de notre rapport.

1.1 Contexte problématique

Les progrès technologiques et la croissance exponentielle du volume de données générées quotidiennement dans divers secteurs ont engendré des défis majeurs en matière de stockage, de traitement et d'analyse de ces données. Les méthodes d'analyse traditionnelles s'avèrent souvent insuffisantes pour exploiter pleinement la valeur cachée au sein de ces ensembles massifs de données. C'est à ce point que l'intelligence artificielle (IA) entre en jeu.

L'IA, en particulier le domaine du machine learning (ML), offre des outils puissants et de nouvelles opportunités pour l'analyse des Big Data. Elle permet de découvrir des schémas complexes, de réaliser des prédictions précises et de prendre des décisions éclairées à partir de ces vastes ensembles de données. Toutefois, malgré les avancées significatives réalisées dans ce domaine, les techniques actuelles d'analyse de Big Data basées sur l'IA peuvent encore être améliorées en termes de performance, de précision et de temps de traitement, entre autres. C'est pourquoi de nouvelles techniques et approches sont nécessaires.

Ainsi, la question qui se pose est la suivante : Comment les avancées de l'IA, en particulier dans le domaine du ML, peuvent-elles contribuer à améliorer et évaluer les techniques d'analyse de Big Data en termes de performance, de précision et de temps de traitement ?

1.2 Objectifs Contribution

Ce mémoire présente une méthode novatrice d'analyse du Big Data, spécifiquement axée sur l'amélioration de la détection du cancer du sein. Le Big Data englobe un ensemble massif de données générées quotidiennement à partir de multiples sources, et il est impératif d'identifier précocement le cancer du sein afin d'accroître les chances de guérison et d'améliorer les résultats cliniques.

L'utilisation du deep learning, plus précisément des réseaux de neurones convolutionnels (CNN), offre de vastes possibilités pour extraire des informations hautement significatives à partir d'images médicales telles que les mammographies. Cette approche permet d'automatiser le processus de détection et de classification des anomalies, contribuant ainsi à réduire les erreurs humaines et à accélérer le processus de diagnostic.

Notre mémoire se concentre sur l'exploration des différentes étapes du traitement du Big Data afin de développer une méthode générale, adaptable à l'analyse de vastes ensembles de données médicales. Nous mettons en évidence l'utilisation des CNN pour extraire des caractéristiques pertinentes à partir des images de mammographie et pour effectuer une classification précise des tissus mammaires, permettant ainsi une évaluation précise des résultats diagnostiques.

Une étude de cas approfondie est présentée afin d'évaluer l'efficacité de la méthode proposée. Les résultats obtenus démontrent de manière significative une amélioration de la détection du cancer du sein par rapport aux méthodes traditionnelles utilisées. Cette approche novatrice ouvre des perspectives prometteuses pour le dépistage précoce et précis du cancer du sein, ce qui peut avoir un impact positif sur le pronostic et la survie des patients. Par conséquent, notre mémoire contribue à l'avancement des connaissances et des pratiques dans le domaine de la détection précoce du cancer du sein en utilisant des techniques d'analyse du Big Data basées sur le deep learning.

1.3 Organisation du mémoire

Le mémoire est structuré en trois chapitres selon le plan suivant, où le chapitre 1 est consacré à l'introduction générale et le chapitre 5 à la conclusion générale :

- **Chapitre 2 : Généralités** : Ce chapitre offre une présentation concise des concepts fondamentaux du Big Data et de l'apprentissage automatique (Machine Learning), en clarifiant leur définition et leur développement au fil du temps. Il examine les caractéristiques essentielles du Big Data ainsi que les divers types d'apprentissage automatique.
- **Chapitre 3 : État de l'art** : Ce chapitre présente une étude approfondie des recherches existantes et des avancées récentes dans le domaine de l'analyse du Big Data utilisant les techniques d'apprentissage automatique (Machine Learning).
- **Chapitre 4 : Contribution et implémentation** : Ce chapitre se focalise sur le traitement du Big Data et se base sur une étude de cas spécifique : l'utilisation de techniques de Deep Learning, plus précisément les réseaux de neurones convolutionnels (CNN).

Chapitre 2

Généralités

2.1 Introduction

La data science est une branche de l'intelligence artificielle qui englobe les domaines interconnectés de la statistique, des méthodes scientifiques et de l'analyse des données. Ces éléments sont utilisés pour extraire du sens et des perspectives à partir des données. Avec l'avènement d'Internet, le monde est devenu extrêmement connecté, où chaque objet manipulé (voitures, réfrigérateurs, vêtements, réseaux sociaux, etc.) génère quotidiennement des millions de données supplémentaires qui s'ajoutent à un vaste océan de données. Toutes ces données peuvent être exploitées pour offrir des services personnalisés et immédiats, répondant ainsi aux attentes des utilisateurs. Mais comment pouvons-nous transformer cet océan de données apparemment infini en un flux régulier d'informations pertinentes ? La réponse réside dans l'intelligence artificielle.

L'intelligence artificielle (IA), ou AI en anglais pour "Artificial Intelligence", englobe un ensemble de techniques visant à permettre aux machines de simuler une forme d'intelligence semblable à celle des êtres humains. L'IA est appliquée dans divers domaines tels que la médecine, le transport, la photographie, etc., avec pour objectif le développement de machines capables de comportements intelligents.

Cependant, un problème se pose actuellement : la croissance exponentielle des données dépasse les capacités des bases de données traditionnelles. Ceci est dû au volume considérable des données, à leur diversité et au temps de traitement qui doit rester dans des délais acceptables

pour les utilisateurs. Les développeurs sont donc confrontés au défi de mettre au point des technologies capables de traiter d'énormes quantités de données variées en un temps réduit, ce que l'on appelle le Big Data.

La **Figure 2.1** illustre l'interrelation entre l'intelligence artificielle (IA), la science des données (Data Science) et le Big Data, qui joue un rôle essentiel dans l'adoption de la transformation numérique. L'IA utilise des algorithmes complexes pour traiter les données, la science des données collecte, traite et analyse les données pour alimenter l'IA, tandis que le Big Data fournit les quantités massives de données nécessaires à l'IA.

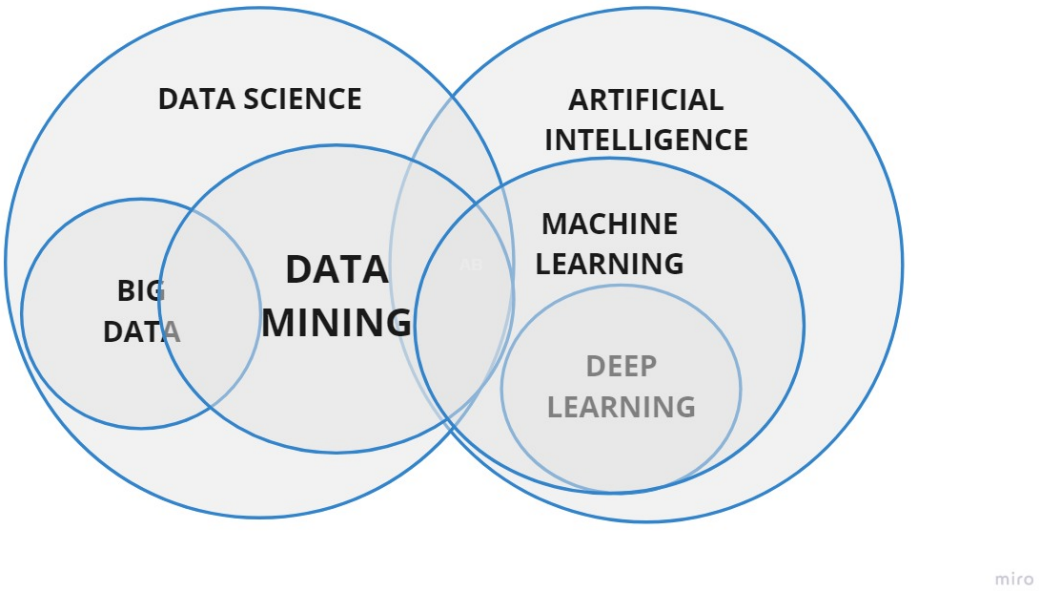


FIGURE 2.1 – Interaction entre l'Intelligence Artificielle, la data science et le big data

Dans la continuation de ce chapitre, nous allons introduire les notions essentielles liées au domaine du "Big Data" et examiner également les concepts de l'apprentissage automatique (Machine Learning).

2.2 Le Big Data et son analyse

Le monde est constamment alimenté par des données et fait l'objet d'une analyse perpétuelle. Le domaine de l'analyse des données (DAD) joue un rôle essentiel dans tous les secteurs en permettant d'extraire du sens des données collectées, ouvrant ainsi la voie à un avenir

extraordinaire. Par exemple, il permet la conception de voitures autonomes sécurisées, le développement de médicaments efficaces et l'amélioration de nos prises de décision grâce à des machines intelligentes, etc.

Bien que l'acronyme de l'analyse des données (ADD) puisse différer de celui du Big Data, il est la clé pour donner du sens à toutes les informations que nous collectons.

2.2.1 Définition du Big Data

Le concept de "Big Data" a suscité plusieurs définitions qui n'ont pas été universellement adoptées en raison de sa complexité et de sa variation selon les utilisateurs et les fournisseurs de services impliqués. Parmi ces définitions, nous pouvons citer :

Contrairement aux données traditionnelles, le terme "Big Data" fait référence à des ensembles de données volumineux et en constante croissance, comprenant des formats hétérogènes tels que des données structurées, non structurées et semi-structurées. Les mégadonnées ont une nature complexe qui nécessite des techniques puissantes et des algorithmes avancés. Ainsi, les outils traditionnels de business intelligence statique ne sont plus efficaces dans le contexte des applications Big Data [8].

Le Big Data, également connu sous le nom de mégadonnées, englobe un ensemble de technologies, d'architectures et de procédures qui permettent d'analyser et de traiter d'importantes quantités de données hétérogènes, afin d'extraire des informations pertinentes à un coût raisonnable [9].

Le terme "Big Data", traduit littéralement par "grosses données" ou "données massives", fait référence à l'explosion de données. On peut également le comparer à la notion de "datamasse" en faisant une analogie avec la biomasse, qui représente un écosystème complexe à grande échelle.

Une autre définition couramment utilisée est celle proposée par IBM[1] :

Le Big Data fait référence aux quantités exponentielles de données, à la fois structurées et non structurées, qui sont si vastes et complexes qu'elles dépassent les capacités des systèmes traditionnels d'information pour les collecter, les stocker, les gérer et les analyser efficacement

afin d'en tirer des informations précieuses.

2.2.2 Évolution historique du Big Data

L'évolution du Big Data peut être divisée en différentes périodes, marquées par des avancées technologiques majeures et des changements dans l'utilisation des données par les entreprises.

- **Les années 1960 et 1970 :** Pendant cette période, les premières bases de données et les premiers systèmes de gestion de bases de données (SGBD) ont été développés. Bien que coûteux et complexes, ces systèmes ont permis aux entreprises de stocker et de traiter de grandes quantités de données.
- **Les années 1970-2000 :** Avec l'augmentation des volumes de données, les entreprises ont commencé à utiliser des SGBD pour stocker et gérer leurs données. Cependant, ces systèmes étaient limités en termes de capacité et de vitesse de traitement.
- **Au début des années 2000 :** Avec l'avènement d'Internet et le développement des technologies de l'information, les entreprises ont commencé à collecter de plus en plus de données en ligne, telles que les données de navigation Web, les données de vente et les données marketing. Cependant, les systèmes informatiques existants étaient incapables de traiter ces quantités massives de données, ce qui a nécessité le développement de nouvelles technologies [10].
- **En 2004 :** Google a publié un article scientifique sur la méthode MapReduce, utilisée pour le traitement des données à grande échelle sur des clusters informatiques [11].
- **En 2006 :** Doug Cutting et Mike Cafarella ont développé le système de gestion de données distribué Hadoop, qui a permis aux entreprises de collecter, stocker et analyser d'énormes volumes de données de manière efficace.
- **Les années 2010-2015 :** L'expression "Big Data" est apparue pour décrire le défi croissant de la gestion des grandes quantités de données. Les entreprises ont adopté des technologies telles que le stockage distribué et le traitement parallèle pour traiter les données massives.
- **Au cours des années 2015 :** Les entreprises continuent d'adopter des solutions Big Data pour collecter, stocker, gérer et analyser les données, en utilisant des technolo-

1. source Site office du BIM : https ://www.ibm.com/cloud/learn/big-data

gies telles que le machine learning et l'IA [9]. Les défis du Big Data comprennent la qualité des données, la sécurité des données et l'analyse efficace des données massives. L'impact de la pandémie de COVID-19 a également entraîné une augmentation de la collecte de données pour surveiller les tendances et les effets de la pandémie. En 2021, les entreprises continuent d'investir massivement dans le Big Data, avec une prévision de dépenses mondiales d'environ 274 milliards de dollars pour l'année [9].

Aujourd'hui, le Big Data est un domaine en constante évolution et les entreprises cherchent à exploiter les avantages de l'analyse des données pour améliorer la prise de décision, le marketing et les opérations commerciales.

2.2.3 Caractéristiques du Big Data

Les traits fondamentaux des mégadonnées ou Big Data, illustrés dans la Figure2.2, sont couramment désignés sous le nom des "Cinq V" :

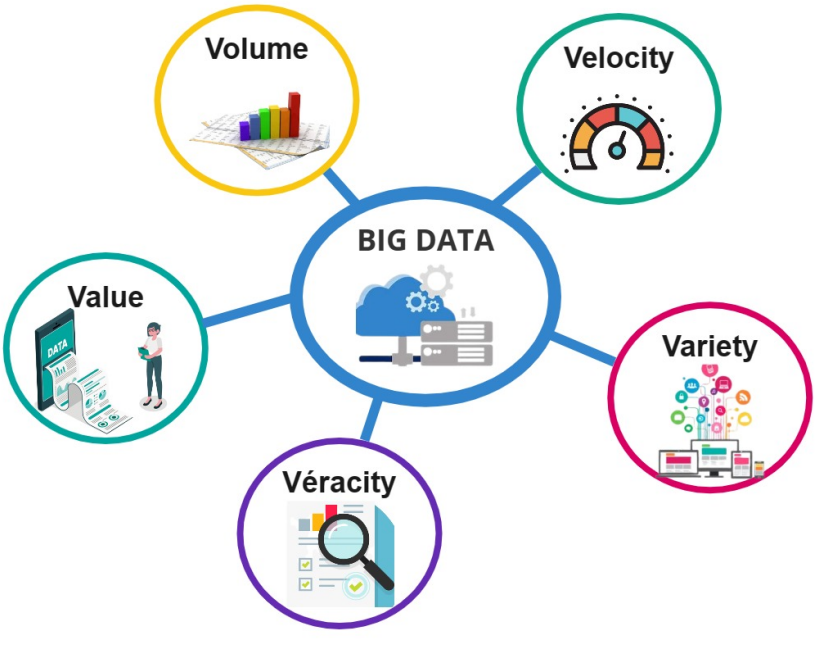


FIGURE 2.2 – Les cinq Vs du Big Data.

Volume :

Le Big Data implique le traitement d'énormes volumes de données non structurées à faible densité. Ces données peuvent provenir de diverses sources, telles que les flux de clics sur des

sites Web, les activités sur les réseaux sociaux ou les mesures de capteurs. Pour certaines organisations, cela peut représenter des dizaines de téraoctets de données, tandis que pour d'autres, cela peut atteindre des centaines de pétaoctets.

Vélocité (Vitesse) :

La vélocité fait référence à la vitesse à laquelle les données sont générées et traitées. Les données à grande vitesse nécessitent une ingestion et un traitement en temps réel. Par exemple, les systèmes de surveillance en temps réel ou les plateformes d'analyse de données en streaming doivent être capables de traiter et de réagir aux données en quasi-temps réel.

Variété :

La variété concerne les différents types de données auxquels le Big Data fait face. Les données traditionnelles étaient principalement structurées et stockées dans des bases de données relationnelles. Cependant, avec l'avènement du Big Data, les données peuvent être non structurées, telles que des données textuelles, audio ou vidéo. Le traitement de ces données nécessite des techniques spécifiques pour les interpréter et les exploiter, ainsi que pour gérer les métadonnées associées.

Véracité :

La véracité se réfère à l'exactitude et à la fiabilité des données. Pour obtenir de la valeur à partir des données, il est essentiel de nettoyer les données pour éliminer les erreurs et les incohérences. La véracité des données est essentielle pour garantir la qualité des analyses et des décisions basées sur ces données.

Valeur :

La valeur représente l'utilité et l'importance des données pour atteindre un objectif spécifique. L'objectif ultime de l'analyse des mégadonnées est d'extraire de la valeur à partir des données collectées. La valeur des données peut également être liée à leur validité et à leur exactitude. Dans certains cas, la valeur dépend également de la rapidité avec laquelle les données peuvent être traitées pour prendre des décisions rapides et éclairées [12].

2.2.4 Structuration du Big Data

La structuration des données est le processus d'organisation et de stockage des données dans un ordinateur de manière à ce qu'elles puissent être référencées et modifiées de manière efficace. Dans le contexte du Big Data, les données collectées, stockées et traitées proviennent de différents domaines et sont générées par de multiples sources de données hétérogènes, ce qui crée une masse de données de natures différentes [13].

2.2.4.1 Données structurées

Les données structurées font référence à des données d'un format et d'une longueur spécifiques, d'une facilité de stockage et d'analyse et d'une organisation élevée. Cela signifie que les données sont organisées dans une structure reconnaissable afin qu'elles puissent répondre aux requêtes pour récupérer des informations à des fins organisationnelles. Une base de données relationnelle comme Structured Query Language (SQL) est un bon exemple de données structurées, elle contient des nombres structurés, des dates, des combinaisons de mots et de nombres appelées chaînes/texte.

En raison de la structure transparente de la base de données, elle peut être recherchée à l'aide d'algorithmes de recherche simples et directs qui peuvent être par type de données dans le contenu réel [14]

2.2.4.2 Données semi-structurées

Les données non structurées sont des informations qui, sous de nombreuses formes différentes, ne correspondent pas aux modèles de données traditionnels et ne conviennent donc généralement pas à une base de données relationnelle traditionnelle. Cela rend le traitement et l'analyse des données non structurées très difficiles et chronophages. Selon Feldman et Sanger, les données non structurées n'ont pas de structure définie.

Les données non structurées incluent généralement des images/objets bitmap, du texte, des e-mails et d'autres types de données qui ne font pas partie de la base de données[15].

Il est important de comprendre la nature et la structure des données afin de déterminer la meilleure approche pour les collecter, les stocker et les analyser. Les solutions Big Data sont capables de gérer différents types de données et de les traiter de manière efficace, permettant

ainsi une analyse plus approfondie et une prise de décision améliorée.

2.2.5 Traitement du Big Data

Le traitement du Big Data comprend généralement les étapes suivantes, illustrées dans la Figure 2.3 :

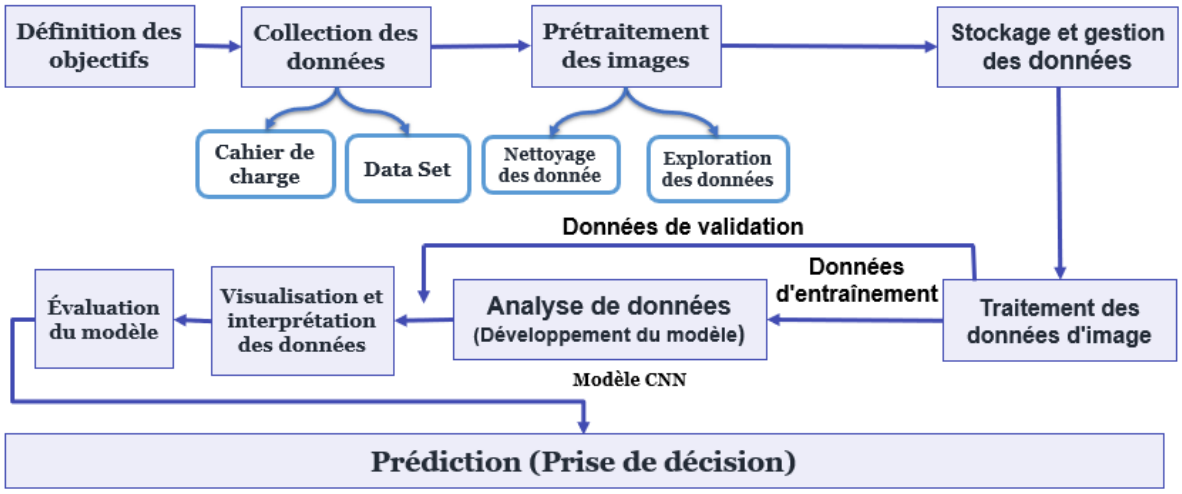


FIGURE 2.3 – Étapes du traitement des Big Data.

2.2.5.1 Définition des objectifs

La première étape du traitement des Big Data consiste à identifier les objectifs à atteindre. Quelles informations ou quelles questions souhaitez-vous obtenir à partir des données ? Il est important de définir clairement les objectifs pour orienter les étapes suivantes.

2.2.5.2 Collecte de données

La collecte de données est la première étape de la gestion des Big Data. Elle implique la récupération de données à partir de différentes sources telles que des bases de données internes, des capteurs connectés, des réseaux sociaux, des fichiers journaux, des données GPS, etc. Ces données peuvent être structurées, semi-structurées ou non structurées. Il est important de collecter suffisamment de données pour obtenir des informations pertinentes, tout en évitant une collecte excessive qui rendrait le processus d'analyse coûteux et complexe [16].

2.2.5.3 Préparation des données

La préparation des données consiste à nettoyer, normaliser et transformer les données pour les rendre utilisables dans le cadre de l'analyse. Cela peut inclure des tâches telles que la suppression des données en double, la correction des erreurs de saisie, la conversion des données non numériques en données numériques, la fusion de données provenant de différentes sources, etc [16].

2.2.5.4 Stockage et gestion des données

Le stockage et la gestion des données font référence aux processus et aux techniques utilisés pour stocker, organiser, sécuriser et gérer les données de manière efficace et fiable. Les Big Data nécessitent souvent des solutions de stockage et de gestion appropriées pour une utilisation efficace. Des solutions telles que les systèmes de fichiers distribués (comme Hadoop HDFS) ou les bases de données NoSQL sont utilisées pour gérer et organiser les données de manière évolutive.

2.2.5.5 Analyse des données

L'analyse des données consiste à utiliser des techniques statistiques et d'apprentissage automatique pour extraire des informations et des insights à partir des données. Des algorithmes d'apprentissage automatique tels que la régression, les arbres de décision, les réseaux neuronaux, etc., peuvent être utilisés pour prédire des tendances, identifier des anomalies, découvrir des relations cachées, etc [16].

2.2.5.6 Visualisation et interprétation des données

La visualisation et l'interprétation des données sont les étapes suivantes de la gestion des Big Data. Cette phase consiste à présenter les résultats de l'analyse sous forme de graphiques, de tableaux et de rapports pour une meilleure compréhension. Des outils de visualisation tels que des graphiques, des tableaux croisés dynamiques et des cartes peuvent être utilisés pour représenter les données de manière claire et concise [16].

2.2.5.7 Intégrité et sécurité des données

La dernière étape de la gestion des Big Data consiste à garantir l'intégrité et la sécurité des données. Il est important de veiller à la qualité, à l'exactitude et à la confidentialité des données en utilisant des méthodes de cryptage, d'authentification et de sauvegarde pour protéger les données contre les risques de sécurité tels que les violations de données, les cyberattaques et les erreurs humaines. Des politiques de sécurité des données doivent être mises en place pour gérer l'accès aux données, les mots de passe et les autorisations d'accès. Il est également important de surveiller régulièrement les données afin de détecter les anomalies et les incohérences [16].

2.3 Apprentissage automatique (Machine Learning)

L'apprentissage automatique, également connu sous le nom de Machine Learning, est une méthode d'analyse de données qui automatise la construction de modèles analytiques. C'est une branche de l'intelligence artificielle (IA) qui repose sur l'idée que les systèmes peuvent apprendre à partir de données, identifier des motifs et prendre des décisions avec un minimum d'intervention humaine. Les algorithmes d'apprentissage automatique peuvent être utilisés pour diverses tâches telles que la reconnaissance d'images et de la parole, le traitement du langage naturel et la prise de décision [8].

Dans cette section, nous examinerons la définition de l'apprentissage automatique et comment elle a évolué au fil du temps, ainsi que quelques types courants du Machine Learning .

2.3.1 Définitions de l'apprentissage automatique

Au fil du temps, la définition de l'apprentissage automatique a évolué pour inclure de nouveaux développements et concepts. Voici quelques-unes de ces définitions :

- En 1997, Tom Mitchell, un chercheur américain, a défini l'apprentissage automatique comme l'étude d'algorithmes informatiques conçus pour effectuer des tâches sans être explicitement programmés pour les accomplir [**?**].
- L'apprentissage automatique est une approche permettant à un ordinateur d'apprendre à effectuer des calculs sans être explicitement programmé. Cette définition a été proposée par Arthur Samuel, un mathématicien américain qui a développé un programme capable

d'apprendre à jouer aux dames en 1959 (Figure 2.4) [1].

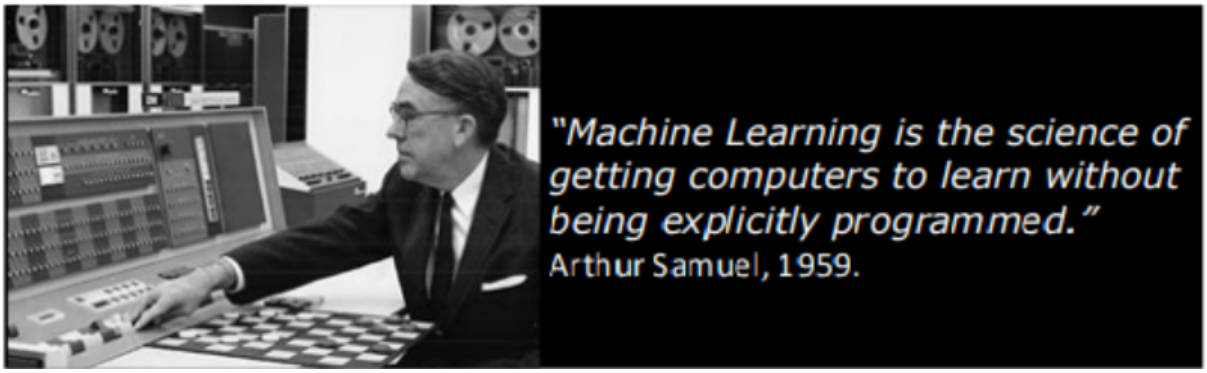


FIGURE 2.4 – Définition de l'apprentissage automatique par Arthur Samuel [1]

- Une définition plus récente, issue d'un sous-ensemble de l'Union européenne en 2020, décrit l'apprentissage automatique comme une approche informatique visant à développer des modèles capables de résoudre des problèmes complexes à partir de données.
  Ces définitions montrent que l'apprentissage automatique est un domaine interdisciplinaire axé sur la création de systèmes informatiques capables d'apprendre de manière autonome à partir de données et de les utiliser pour résoudre des problèmes complexes. Les algorithmes d'apprentissage automatique sont utilisés dans une grande variété d'applications telles que la médecine, le filtrage des e-mails, la reconnaissance vocale et la vision par ordinateur.

2.3.2 Évolution de l'apprentissage automatique

L'histoire de l'apprentissage automatique est riche et remonte à plusieurs décennies. Elle a été influencée par les développements clés dans les domaines de l'intelligence artificielle et de la statistique.
Au début des années 1950, les premiers concepts d'apprentissage automatique ont été proposés par des chercheurs en IA tels que Marvin Minsky et John McCarthy. Dans les années 1960, les algorithmes de régression linéaire et les arbres de décision ont été introduits pour effectuer des prédictions à partir de données. Les algorithmes d'apprentissage supervisé, tels que les réseaux de neurones, sont apparus dans les années 1980 pour traiter des problèmes complexes. Au début des années 1990, les algorithmes d'apprentissage non supervisé, tels que les algorithmes de regroupement (clustering), ont été introduits pour explorer les structures cachées des données. Les algorithmes de Deep

Learning, qui utilisent des réseaux de neurones profonds, ont été développés dans les années 2000 pour traiter des données complexes telles que les images et les textes. Dans les années 2010, les algorithmes d'apprentissage par renforcement, qui permettent aux machines de s'adapter à leur environnement en apprenant à partir de retours d'information, ont été introduits. Plus récemment, les algorithmes d'apprentissage automatique distribué, en temps réel et pour la prise de décision et l'optimisation, sont devenus de plus en plus populaires pour résoudre des problèmes complexes en temps réel.
En résumé, l'histoire de l'apprentissage automatique montre comment ce domaine est devenu un élément clé de l'IA et comment les algorithmes ont évolué pour traiter des données de plus en plus complexes et des problèmes plus difficiles.

2.3.3 Types de Machine Learning

2.3.3.1 Apprentissage supervisé

L'apprentissage supervisé est une méthode d'apprentissage automatique dans laquelle un algorithme informatique est formé à partir d'exemples de données étiquetées pour effectuer une tâche spécifique. Dans un système d'apprentissage supervisé, les algorithmes utilisent les données d'entraînement pour apprendre à faire des prédictions sur de nouvelles données. Les algorithmes apprennent en comparant leurs prédictions à des étiquettes correctes pour les données d'entraînement et en ajustant leurs modèles en conséquence.
L'apprentissage supervisé est utilisé pour de nombreuses tâches, telles que la classification de données, la régression linéaire, la reconnaissance d'images, la reconnaissance de la parole, etc. Il est largement utilisé dans les domaines de la science des données, de l'intelligence artificielle (IA) et du machine learning (ML) [10].

2.3.3.2 Apprentissage non supervisé

L'apprentissage non supervisé (Unsupervised Learning) est une méthode du ML dans laquelle un algorithme informatique est formé à partir de données non étiquetées pour découvrir des structures et des relations dans les données. Dans un système d'apprentissage non supervisé, l'algorithme est libre de découvrir des motifs et des modèles dans les données sans guidance explicite. Les algorithmes peuvent utiliser des techniques

telles que la réduction de dimensions, la clusterisation et la génération de modèles pour découvrir des structures cachées dans les données (voir Figure 2.5).

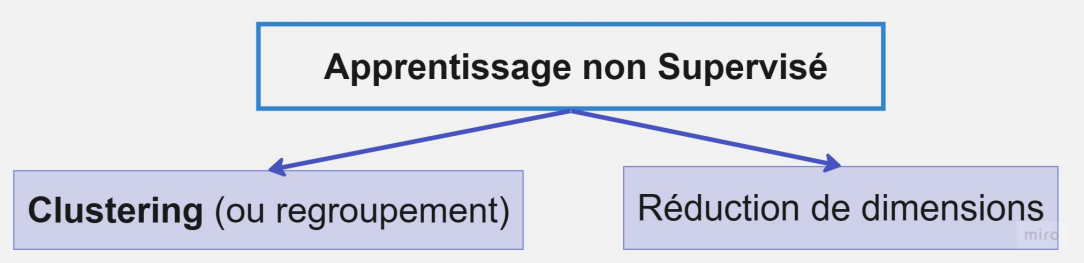


FIGURE 2.5 – Apprentissage non supervisé.

L'algorithme des k-moyennes et les algorithmes Apriori sont des exemples d'apprentissage non supervisé [10].

2.3.3.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une méthode de machine learning qui combine des données étiquetées et non étiquetées pour entraîner un algorithme. Cette approche permet à l'algorithme d'avoir accès à une petite quantité de données étiquetées, qui peuvent fournir des informations précieuses pour guider le processus d'apprentissage, ainsi qu'à une grande quantité de données non étiquetées, qui peuvent améliorer la capacité de généralisation du modèle.
L'apprentissage semi-supervisé se situe entre l'apprentissage supervisé classique, où tous les échantillons sont étiquetés, et l'apprentissage non supervisé, où aucune classe n'est attribuée. Les méthodes d'apprentissage semi-supervisé étendent les techniques d'apprentissage choisi, soit non supervisé ou supervisé, pour inclure des informations supplémentaires typiques de l'autre paradigme d'apprentissage.

2.3.3.4 Apprentissage par renforcement

L'apprentissage par renforcement fait référence à une classe de problèmes d'apprentissage automatique dans lesquels les machines learning essa
ient différentes situations afin de pouvoir déterminer quelles actions sont les plus utiles, et pas seulement recevoir des instructions sur les actions à appliquer, ce qui distingue cette méthode des autres techniques d'apprentissage.
Il est également considéré comme une forme d'apprentissage comportemental. L'algo-

rithme dans ce cas reçoit les informations en analysant les données pour pouvoir orienter l'utilisateur vers les meilleurs résultats. Dans ce type d'apprentissage, le système n'est pas entraîné à partir d'un ensemble de données, mais apprend par l'expérience et utilise les erreurs, par exemple pour les voitures autonomes, ce qui diffère des autres types d'apprentissage supervisé [17].

2.3.3.5 Deep Learning

Le deep learning (DL) est un type de machine learning qui utilise des réseaux de neurones artificiels pour modéliser et résoudre des problèmes complexes. Ces réseaux de neurones sont composés de couches de nœuds interconnectés, ou "neurones" (comme illustré dans la Figure 2.6), conçus pour traiter et analyser de grandes quantités de données d'entrée. Chaque neurone dans un réseau de deep learning reçoit une entrée des neurones de la couche précédente et utilise cette entrée pour effectuer une opération mathématique, appelée fonction d'activation, qui produit une sortie. Cette sortie est ensuite transmise à la couche suivante de neurones, et le processus est répété jusqu'à ce que la sortie finale soit produite.

Les algorithmes de deep learning peuvent être formés de manière supervisée ou non supervisée, en fonction du type de problème à résoudre.

L'apprentissage supervisé consiste à fournir à l'algorithme des données d'entraînement étiquetées, où la sortie correcte ou "étiquette" est connue pour chaque entrée. L'algorithme apprend alors à faire des prévisions sur de nouvelles données non vues en se basant sur les modèles qu'il a identifiés dans les données d'entraînement.

L'apprentissage non supervisé, d'autre part, consiste à fournir à l'algorithme des données non étiquetées, et l'algorithme doit identifier des modèles ou des caractéristiques dans les données par lui-même.

Le deep learning est particulièrement adapté aux tâches qui impliquent la reconnaissance d'images et de la parole, le traitement du langage naturel et d'autres types de données non structurées. En raison de sa capacité à apprendre des modèles complexes, il obtient des résultats de pointe dans de nombreux domaines. Cependant, il nécessite également des ressources informatiques importantes et de grandes quantités de données étiquetées.

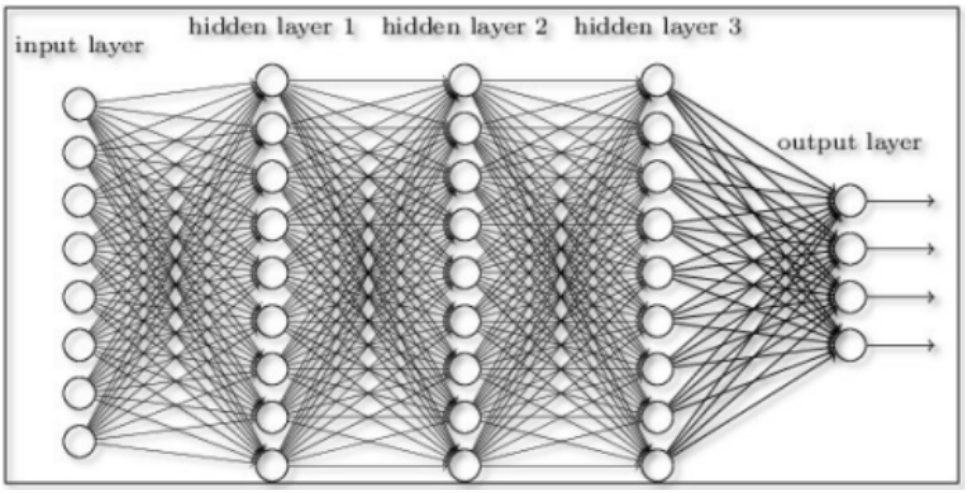


FIGURE 2.6 – Deep learning (réseaux neuronaux). [1]

2.4 Conclusion

En conclusion de ce chapitre, nous avons exploré les concepts essentiels du Big Data et du Machine Learning (ML). Nous avons défini le Big Data comme étant un ensemble de données volumineux et complexes, et nous avons également montré comment il a évolué au fil du temps. De plus, nous avons examiné les caractéristiques clés du Big Data, telles que la variété, la vélocité et la véracité, et nous avons discuté de la structuration et de la gestion de ces données massives, ainsi que de leur importance pour extraire des informations utiles.

En ce qui concerne le Machine Learning, nous avons défini cette technologie comme un sous-domaine de l'intelligence artificielle qui permet aux ordinateurs d'apprendre sans être explicitement programmés. Nous avons examiné les différents types de Machine Learning, y compris l'apprentissage supervisé, non supervisé, semi-supervisé, renforcé et Deep Learning.

En résumé, ce chapitre met en évidence le rôle crucial du Big Data et du Machine Learning dans l'analyse des données massives et la résolution de problèmes complexes. Leur popularité ne cesse de croître et nous pouvons nous attendre à ce que leur utilisation continue de se développer alors que les entreprises cherchent à exploiter pleinement le potentiel de la transformation numérique.

Dans le chapitre suivant, nous nous concentrerons sur les avancées récentes de la recherche dans les techniques de Machine Learning appliquées à l'analyse du Big Data.

Chapitre 3

État de l'art

3.1 Introduction

Dans ce chapitre, nous avons réalisé une étude approfondie sur les avancées récentes de la recherche concernant les techniques de Machine Learning utilisées pour l'analyse du Big Data. Ces avancées nous ont permis de proposer une nouvelle technique d'analyse du Big Data, qui sera présentée en détail dans le chapitre suivant.

3.2 IA au service de l'analyse du Big Data : Classification des techniques et paramètres d'évaluation

L'analyse du Big Data est une entreprise complexe qui requiert des outils et des techniques appropriés. L'intelligence artificielle (IA) est devenue un élément essentiel de l'analyse du Big Data, offrant des méthodes efficaces pour découvrir des modèles, prédire des tendances et prendre des décisions automatisées. Le Machine Learning, en particulier le Deep Learning, a révolutionné notre capacité à traiter les Big Data.
La **Figure 3.1** présente une **taxonomie des techniques d'analyse de Big Data basée sur les sous-domaines de l'intelligence artificielle**. La classification proposée comprend quatre grandes catégories, à savoir le Machine Learning, les méthodes basées sur les connaissances et les inférences, les algorithmes de prise de décision, les méthodes de recherche et la théorie de l'optimisation **(Russell et Norvig, 2020)[18]**. De plus, les paramètres qualitatifs les plus importants pour évaluer chaque méthode d'analyse du

Big Data et comprendre ses avantages et ses inconvénients sont identifiés comme suit [19] :

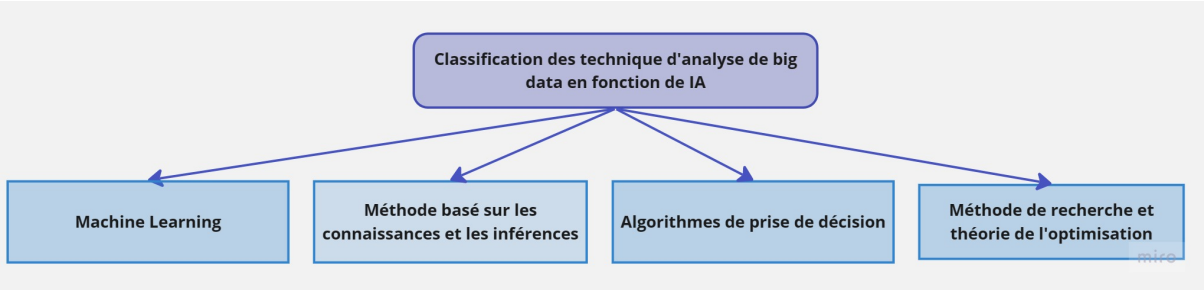


FIGURE 3.1 – Classification des techniques d'analyse big data en utilsant l'IA.

— **Scalabilité :** Il s'agit de la capacité d'un mécanisme ou d'un système à s'adapter rapidement aux changements, tels qu'une augmentation de la taille des données ou du nombre d'utilisateurs, sans compromettre la qualité de l'analyse.
— **Efficacité :** Cela fait référence au rapport entre la méthode utilisée et les besoins totaux en termes de temps et de coûts.
— **Précision :** Elle est évaluée en tenant compte de divers paramètres, tels que les erreurs de données et la capacité prédictive des algorithmes, pour déterminer la fiabilité et l'exactitude des résultats.
— **Confidentialité :** Cela concerne les pratiques mises en place pour garantir que les données sont utilisées uniquement à des fins autorisées et qu'elles sont protégées contre l'accès non autorisé.
— **Complexité :** La complexité d'un algorithme mesure la quantité de ressources nécessaires pour l'exécuter en fonction de la taille de l'entrée. Elle est généralement exprimée en notation Big O et permet de choisir les algorithmes les plus efficaces pour résoudre un problème donné.

Il est important que ces concepts sont cruciaux lors du développement et de l'application de nouvelles techniques d'analyse de Big Data basées sur l'IA. Les chercheurs et les praticiens doivent prendre en compte ces aspects pour assurer des résultats précis, une efficacité opérationnelle, une évolutivité et une confidentialité appropriées dans leurs solutions d'analyse des Big Data.

3.3 Mécanismes du Machine Learning

Il existe plusieurs méthodes de Machine Learning utilisées pour l'analyse du Big Data, chacune ayant ses avantages et ses limites. La **Figure 3.2** présente certains des algorithmes les plus couramment utilisés pour chaque type de Machine Learning :

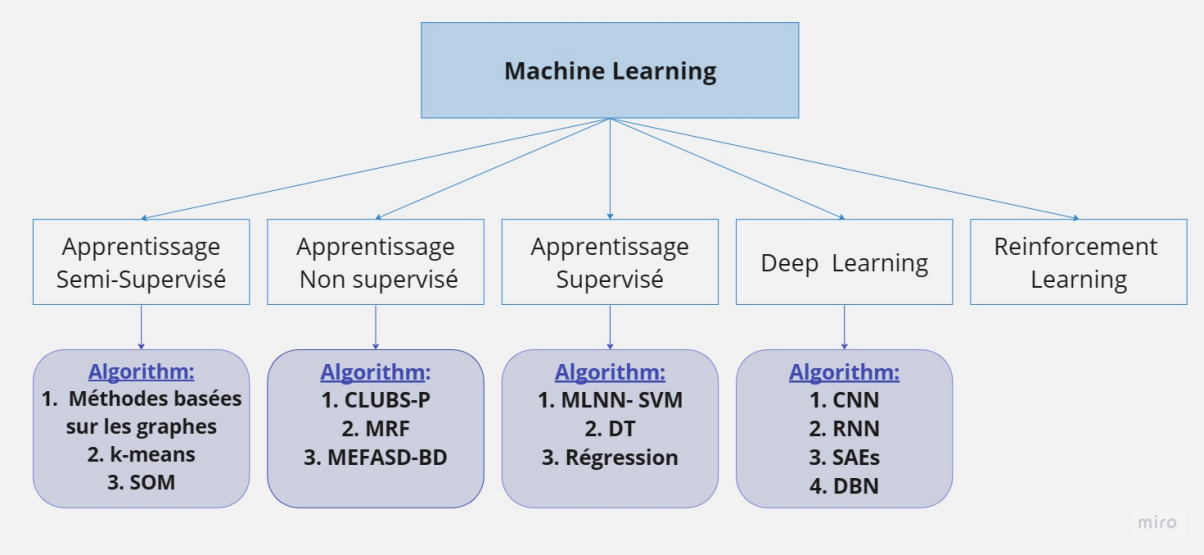


FIGURE 3.2 – Taxonomie des algorithmes les plus courants pour chaque type de Machine Learning utilisé pour l'analyse du Big Data.

3.3.1 Apprentissage supervisé

La première catégorie nécessite des efforts manuels considérables pour mettre les données dans un format adapté aux algorithmes d'apprentissage. Il existe plusieurs algorithmes de Machine Learning supervisés couramment utilisés pour l'analyse de gros volumes de données (Big Data). Ces algorithmes sont souvent utilisés pour résoudre des problèmes de classification, où le modèle apprend à prédire des étiquettes discrètes ou des catégories prédéfinies pour de nouvelles données, ou de régression, où le modèle apprend à prédire des valeurs continues ou numériques pour de nouvelles données (voir la **Figure 3.3**), qui sont des tâches courantes dans l'analyse de données. Voici quelques-uns des algorithmes les plus couramment utilisés :

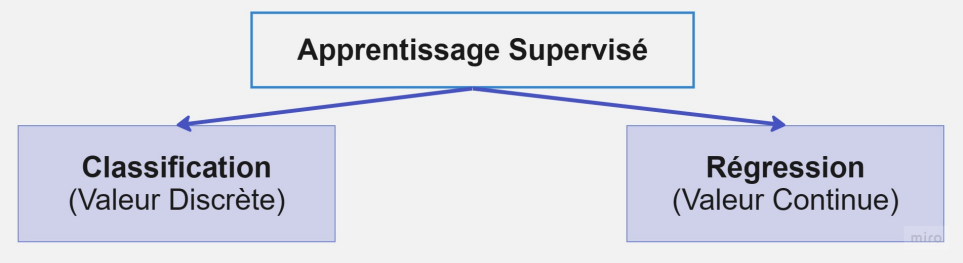


FIGURE 3.3 – L'apprentissage supervisé

3.3.1.1 Réseau neuronal multicouche entraîné avec Support Vector Machine (SVM)

Dans leur article, **AlZubi (2020) [20]**, les auteurs ont développé une nouvelle technique de classification de données massives pour la reconnaissance du diabète en utilisant des techniques de MapReduce et de Machine Learning. Ils collectent d'abord les données à partir du Big Data et utilisent le modèle MapReduce pour diviser efficacement les ensembles de données en sous-ensembles plus petits. Ensuite, une procédure de nettoyage des données est utilisée pour éliminer le bruit généré par les données collectées. Les caractéristiques sélectionnées sont ensuite formées à l'aide d'un réseau neuronal multicouche entraîné avec SVM. Le réseau neuronal est appliqué pour classer les nouveaux échantillons. Les résultats expérimentaux montrent que cette méthode est efficace pour la reconnaissance du diabète, avec des taux d'erreur, de sensibilité, de spécificité et d'exactitude acceptables.

3.3.1.2 Modèle d'arbre de décision (DT)

Dans leur article, **Nair et Shetty (2018) [2]**, les auteurs ont développé un système de prédiction en temps réel de l'état de santé à distance en appliquant des modèles de Machine Learning à d'énormes flux de données. Le système est conçu en utilisant Apache Spark dans un environnement cloud. Un modèle d'arbre de décision est créé à partir des données de santé actuelles et appliqué au flux de données pour
prédire l'état de santé. Le modèle d'arbre de décision joue un rôle très important dans le Machine Learning. Il est capable de gérer à la fois les variables continues et confidentielles, et fournit une indication claire de prédiction ou de classification sans nécessiter beaucoup de calculs, permettant ainsi une meilleure approximation, quelle que soit la complexité des données. Les algorithmes les plus couramment utilisés pour cela sont "Classification and Regression Trees" (CART) et "Iterative Dichotomiser 3" (ID3).

La structure présentée conduit à une efficacité optimale en termes de temps et de système. La confidentialité des données est protégée en utilisant un compte secondaire Twitter.

La **Figure 3.4** suivante représente ce système évolutif : l'utilisateur tweete ses attributs de santé, et l'application les reçoit en temps réel, extrait les attributs et applique le modèle de Machine Learning pour prédire son état de santé. Les résultats sont ensuite envoyés directement à l'utilisateur via un message instantané pour prendre les mesures appropriées.

En conclusion, le modèle d'arbre de décision permet d'économiser considérablement de temps et d'argent en utilisant pleinement les technologies existantes plutôt que de développer de nouvelles technologies pour effectuer la même tâche. Avec quelques modifications mineures, cette application peut être utilisée pour prédire la présence de diverses maladies.

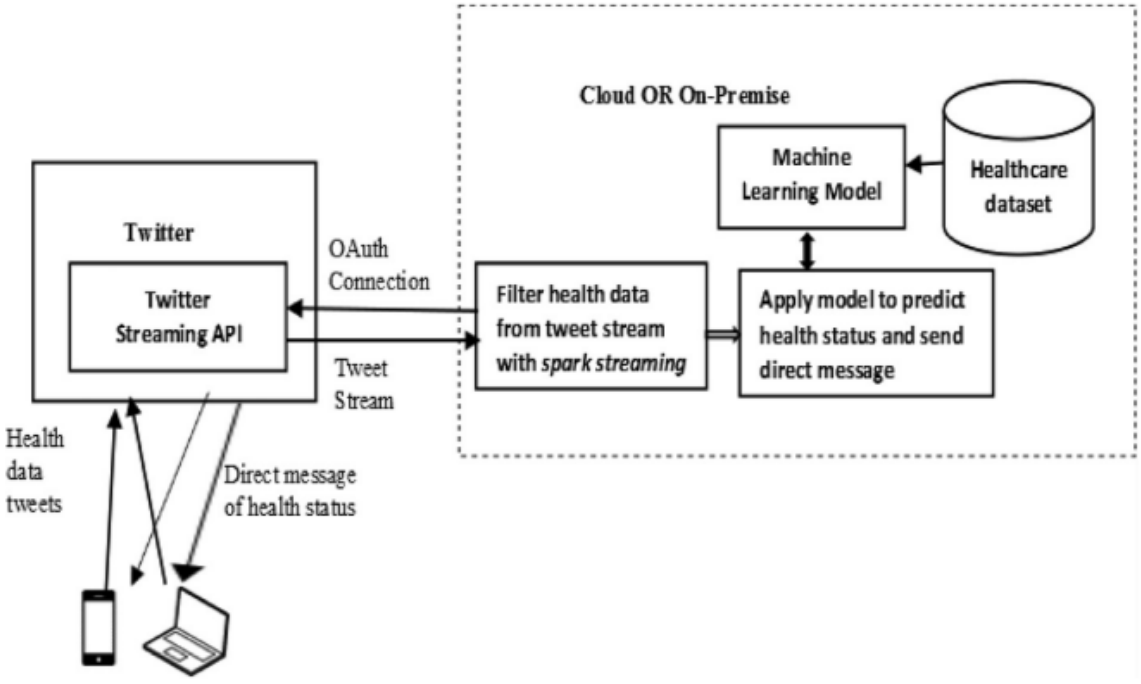


FIGURE 3.4 – Modèle de système de surveillance de la santé en temps réel [2].

3.3.1.3 Algorithmes de régression avec apprentissage d'ensemble

Dans une étude menée par **Asencio-Cortés et al. (2018) [3]**, l'objectif était de prédire les événements sismiques en Californie en utilisant des algorithmes de régression combinés à un apprentissage d'ensemble, dans le contexte du Big Data. Les chercheurs ont

analysé les données sur les tremblements de terre en Californie de 1970 à 2017, totalisant 1 Go d'informations réparties en 27 ensembles de données. L'objectif était de prédire la magnitude des tremblements de terre pour les sept prochains jours.
Leur approche utilisait le framework Apache Spark ainsi que des modèles d'apprentissage automatique de la bibliothèque H2O, tels que la régression linéaire, les machines à renforcement de gradient (GBM), l'apprentissage profond et les forêts aléatoires (voir **Figure 3.5**). Ils ont également exploité l'infrastructure cloud d'Amazon. Les résultats obtenus étaient très prometteurs, avec des erreurs relatives avoisinant les 10
Les principaux avantages de la méthode proposée étaient des niveaux élevés de parallélisme et de mise à l'échelle. Cependant, elle présentait une efficacité relativement faible pour le traitement de grands ensembles de données.
En résumé, l'analyse de données massives pour la prédiction de la magnitude des séismes ouvre une nouvelle voie de recherche très prometteuse. Cette approche pourrait être utile pour traiter simultanément de grandes quantités de données contenant de nombreuses variables.

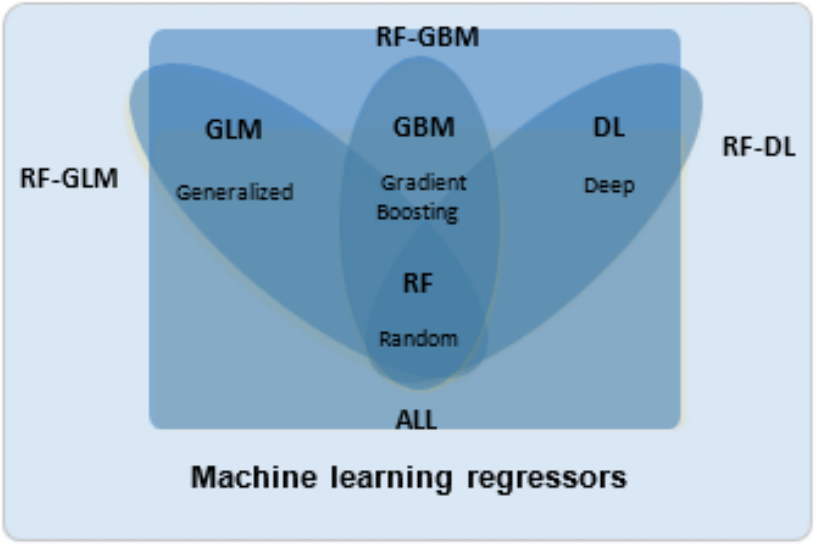


FIGURE 3.5 – Exemple d'algorithmes de régression avec apprentissage d'ensemble [3].

Le **Tableau 3.1** fourni présente une comparaison entre différents algorithmes d'apprentissage supervisé :

TABLE 3.1 – Comparaison entre les algorithmes supervised learning [6].

Références	Algorithme	Évolutivité	Efficacité	Précision	Confidentialité
(BDA-MC)[20]	**MLNN entraîné avec SVM**	X	X	oui	X
ML-SBD [2]	**Le modèle d'arbre de décision**	oui	oui	oui	oui
RAC-BD [3]	**régression**	oui	X	oui	X

3.3.2 Apprentissage non supervisé

Les algorithmes d'apprentissage non supervisé sont utilisés pour découvrir des structures ou des modèles cachés dans les données sans avoir besoin d'une étiquette ou d'une réponse prédéfinie pour chaque observation. Un exemple courant d'application d'apprentissage non supervisé est l'algorithme de recommandation de produits d'Amazon [21].

Les algorithmes non supervisés peuvent être classés en deux catégories principales : le clustering (ou regroupement), qui regroupe les données similaires en clusters, et la réduction de dimensionnalité, qui permet de visualiser les données pour faciliter leur compréhension.

3.3.2.1 CLUBS-P

Dans leur article, **Ianni et al. (2020) [22]** soulignent que les méthodes traditionnelles d'analyse de données, telles que l'analyse de régression et les tests d'hypothèses, ne sont pas adaptées à la gestion des big data générées par les applications et les dispositifs modernes. Ils proposent d'utiliser des algorithmes de clustering qui peuvent identifier des modèles et des relations dans les données sans nécessiter de connaissances ou d'hypothèses préalables.

Dans leur étude, **Ianni et al. (2020) [22]** ont présenté un algorithme de clustering basé sur le centroïde pour les big data, appelé CLUBS-P. Ils ont évalué deux implémenta-

tions différentes de CLUBS-P pour déterminer sa faisabilité dans le contexte des big data. Les résultats obtenus ont montré une très bonne évolutivité de l'algorithme et une amélioration par rapport à l'algorithme de clustering parallèle de référence, à savoir K-means. De plus, l'implémentation de CLUBS-P basée sur le passage de messages a permis une utilisation optimale des ressources des nœuds disponibles lorsque toutes les fonctionnalités offertes par Spark n'étaient pas nécessaires.

3.3.2.2 Systèmes flous évolutifs multi-objectifs

L'article **Pulgar-Rubio et al. (2017) [23]** présente une nouvelle méthode appelée MEFASD-BD (Multi-objective Evolutionary Fuzzy Systems) qui est la première approche d'évolution des systèmes flous pour la découverte de sous-groupes dans le contexte du Big Data. Cette méthode utilise le modèle MapReduce avec Apache Spark pour traiter efficacement de grandes quantités de données. L'objectif est d'analyser la qualité des sous-groupes obtenus pour chaque bloc de données afin d'améliorer leur pertinence. MEFASD-BD est spécialement conçu pour gérer des ensembles de données de haute dimension, ce qui en fait une solution adaptée aux défis du Big Data.
Les résultats des tests réalisés ont démontré une réduction significative du temps d'exécution tout en maintenant des niveaux élevés de qualité. Cette méthode ouvre de nouvelles perspectives pour la découverte de sous-groupes dans les environnements de Big Data.

3.3.2.3 Champs aléatoires de Markov (MRF)

Dans l'article **Ip et al. (2018) [24]**, un aperçu du Big Data et des techniques d'apprentissage automatique dans la protection des cultures est présenté. De plus, la possibilité d'utiliser des champs aléatoires de Markov (Markov random fields) qui prennent en compte la composante spatiale entre les sites voisins pour modéliser la résistance aux herbicides est examinée. Les résultats de l'expérience ont démontré la performance de l'approche proposée.
Le Tableau 3.2 présente une comparaison entre trois algorithmes d'apprentissage non supervisé :

TABLE 3.2 – Comparaison entre les algorithmes Unsupervised learning [6]

Références	Algorithme	Évolutivité	Efficacité	Précision	Confidentialité
EBD-C [22]	**CLUBS-P**	oui	oui	oui	X
MEFASD-BD [23]	**Systèmes flous évolutifs multi-objectifs**	oui	oui	oui	X
BDML-CP [24]	**MarKov random fields**	oui	X	oui	X

3.3.3 Apprentissage semi-supervisé

L'analyse de Big Data peut être réalisée à l'aide d'algorithmes d'apprentissage semi-supervisé pour traiter de grandes quantités de données de manière plus efficace. L'idée sous-jacente est que les données non étiquetées contiennent des informations utiles qui peuvent être exploitées pour améliorer la généralisation des modèles. Voici quelques exemples d'algorithmes semi-supervisés qui peuvent être utilisés pour l'analyse de Big Data :

3.3.3.1 Méthodes basées sur les graphes

Les méthodes de classification semi-supervisées basées sur les graphes, présentées par **Zhu (2005) [25]**, définissent un graphe où les nœuds représentent des exemples étiquetés et non étiquetés de l'ensemble de données, et les arêtes reflètent la similarité entre les exemples (elles peuvent être pondérées). Ces méthodes utilisent généralement des techniques de propagation d'étiquettes sur le graphe. Les méthodes basées sur les graphes sont généralement non paramétriques, discriminantes et transformationnelles, ce qui signifie qu'elles ne peuvent pas être facilement étendues à de nouveaux points qui ne font pas partie de l'ensemble d'apprentissage.

3.3.3.2 Méthode des k-means

La méthode des k-means, proposée par **MacQueen (1967) [26]**, est une méthode de clustering utilisée pour diviser automatiquement un ensemble de données en k groupes. Elle fonctionne en sélectionnant k centres de groupe initiaux, puis en les affinant itérativement de la manière suivante : chaque instance est assignée au centre de groupe le

plus proche, puis chaque centre de groupe est mis à jour pour être la moyenne des instances qui lui sont assignées. L'algorithme converge lorsque l'affectation des instances aux clusters ne change plus. Les clusters sont pré-initialisés avec des instances choisies au hasard dans l'ensemble de données.

3.3.3.3 Self Organizing Map (SOM)

Le Self Organizing Map (SOM), présenté par **Haykin (2003) [27]**, est un algorithme populaire de réseau de neurones artificiels basé sur l'apprentissage non supervisé. Le SOM peut projeter des données de grande dimension dans un espace de dimension inférieure, ce qui peut être utile pour analyser des modèles dans l'espace de recherche ayant une structure complexe. Les résultats sont visuels et faciles à analyser. Son rôle principal est de réaliser une projection non linéaire des données de haute dimension sur un espace de faible dimension. Les cartes auto-organisatrices sont largement utilisées dans la classification de données.

Le Tableau suivant présente une comparaison entre trois algorithmes d'apprentissage semi-supervisé :

TABLE 3.3 – Comparaison entre les algorithmes semi-supervised learning.

Références	**Algorithme**	**Évolutivité**	**Efficacité**	**Précision**	**Confidentialité**
MCA-MO[26]	**Méthode des k-means**	X	oui	oui	X
SSL-G [25]	**Méthodes basées sur les graphes**	X	oui	oui	X
NN-ACF[27]	**Self Organizing Map**	oui	oui	oui	oui

3.3.4 Deep Learning

Les techniques de Deep Learning sont un domaine de l'informatique qui se concentre sur le développement de réseaux de neurones artificiels capables d'analyser et d'apprendre à partir de données pour prendre des décisions et faire des prédictions. Les données massives sont l'un des principaux domaines où les techniques de Deep Learning sont appliquées, car la quantité de données générées aujourd'hui est trop importante

pour que les techniques traditionnelles puissent les traiter efficacement.
Dans cette partie, nous présenterons quelques-uns des algorithmes de Deep Learning les plus couramment utilisés pour l'analyse de Big Data :

3.3.4.1 Convolutional Neural Networks

Li Shen et al. (2019) [4] présentent un algorithme de Deep Learning, en particulier les Convolutional Neural Networks (CNN), pour la détection précise du cancer du sein à partir d'images mammographiques de dépistage. L'approche "end-to-end" utilisée permet d'obtenir d'excellentes performances, avec des AUC élevées et une réduction des faux positifs et des faux négatifs.
Cette approche permet de réduire significativement le besoin d'annotations de ROI et a de nombreuses applications en imagerie médicale, en plus de la détection du cancer du sein sur les mammographies de dépistage.
L'algorithme a été testé sur deux ensembles de données indépendants : des mammographies en film numérisé de la base de données numérique pour la mammographie de dépistage (CBIS-DDSM) et des images de mammographie numérique plein champ (FFDM) de la base de données INbreast. Les résultats montrent une performance remarquable avec une AUC (Area Under the Curve) par image de 0,88 pour le meilleur modèle individuel sur l'ensemble CBIS-DDSM, et une amélioration à 0,91 grâce à une moyenne de quatre modèles. Sur l'ensemble INbreast, l'AUC par image atteint 0,95 pour le meilleur modèle individuel et 0,98 avec la moyenne de quatre modèles. Les sensibilité et spécificité des modèles sont également rapportées.
Une découverte importante de l'étude est la possibilité de transférer le classificateur d'image entière entraîné sur les mammographies CBIS-DDSM vers les images FFDM INbreast, en utilisant uniquement un sous-ensemble des données INbreast pour l'affinage, sans nécessiter d'annotations de lésions supplémentaires.
La Figure 3.6 représente l'approche end-to-end des Convolutional Neural Networks pour la détection du cancer du sein.

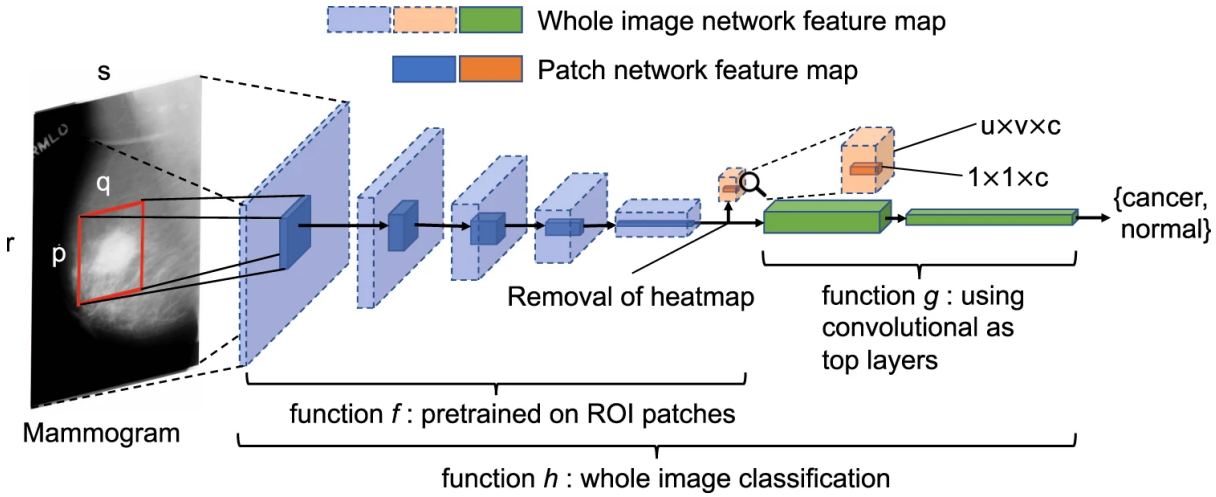


FIGURE 3.6 – Approche end-to-end des Convolutional Neural Networks pour la détection du cancer du sein [4]

3.3.4.2 Recurrent Neural Networks

Dans leur article, Bilal Jan et al. (2017) [28] expliquent que les modèles traditionnels de Deep Learning tels que les autoencodeurs empilés, les Deep Belief Networks (DBN) et les Convolutional Neural Networks (CNN) ne

sont pas adaptés à l'apprentissage de caractéristiques pour les données de séries temporelles. Les Recurrent Neural Networks (RNN) sont présentés comme une solution efficace pour traiter les données de séries temporelles, en particulier dans les applications de traitement du langage naturel. Les RNN apprennent les caractéristiques des données de séries temporelles en stockant l'information des entrées précédentes dans leur état interne, ce qui leur permet de prendre en compte les relations entre les données successives. Cependant, les RNN peuvent rencontrer des problèmes de disparition du gradient, ce qui limite leur capacité à capturer des dépendances à long terme. Des variantes des RNN ont été développées pour résoudre ce problème, telles que les Long Short-Term Memory (LSTM) et les Gated Recurrent Units (GRU). Les RNN ont obtenu des performances remarquables dans de nombreuses applications de traitement du langage naturel, de reconnaissance vocale et de traduction automatique.

La Figure 3.7 représente l'architecture des Recurrent Neural Networks.

Hammou, Lahcen et Mouline (2020) [29] ont introduit une nouvelle technologie efficace pour l'analyse émotionnelle en utilisant des Réseaux de Neurones Récurrents. Les auteurs ont adopté un texte rapide avec des variables de réseau de neurones récurrents

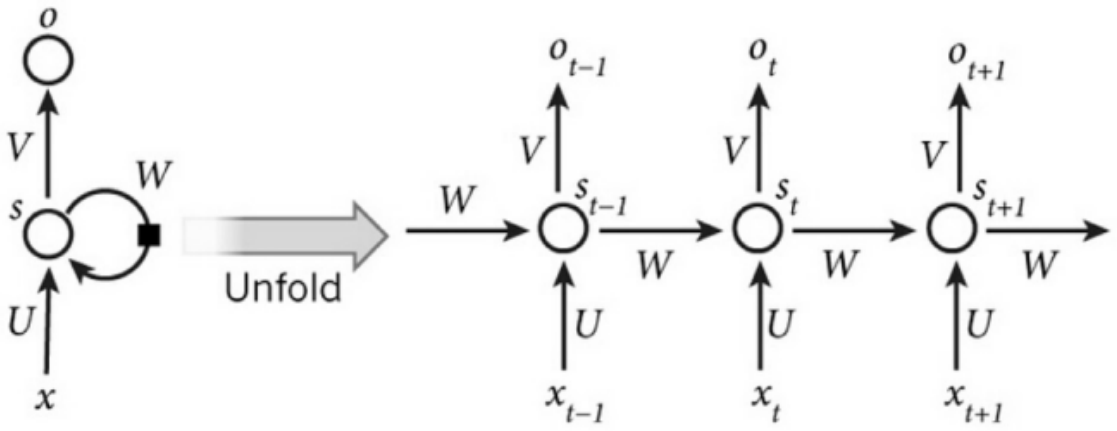


FIGURE 3.7 – Architecture des Recurrent Neural Networks [**?**]

pour représenter et classer les données de texte. De plus, un système distribué basé sur l'apprentissage automatique distribué pour des analyses en temps réel a été suggéré. Les tests effectués démontrent que la méthode fournie dépasse les méthodes LSTM, BTSTM et GRU en termes de précision de la classification, et peut gérer d'excellentes données pour l'analyse émotionnelle.

Les RNN sont particulièrement efficaces lorsqu'il s'agit de traiter de gros volumes de données, car leur capacité à traiter des données séquentielles leur permet d'extraire des caractéristiques et des modèles à partir de vastes quantités de données de manière efficace. De plus, comme les RNN apprennent à partir de données passées, ils peuvent s'adapter aux ensembles de données changeants et aux nouvelles informations, ce qui les rend hautement adaptables et utiles dans diverses applications telles que la reconnaissance de la parole, la classification d'images, la prédiction de séquences et la compréhension du langage naturel. Dans l'ensemble, les RNN sont une technique puissante et polyvalente d'apprentissage en profondeur qui est précieuse pour la manipulation de gros volumes de données.

3.3.4.3 Autoencoders et Stacked Autoencoders (SAEs)

Akrufa Hajirahimova et Aybeniz Aliyeva (2020) [5] présentent les Stacked Autoencoders (SAEs) comme l'une des techniques de Deep Learning les plus largement utilisées. Les SAEs sont construits en empilant plusieurs autoencodeurs, qui

sont les réseaux de neurones feed-forward les plus typiques. Un autoencodeur est une structure d'apprentissage non supervisée qui possède trois couches : la couche d'entrée,

la couche cachée et la couche de sortie, comme représenté dans la Figure 3.8. Le processus d'entraînement d'un autoencodeur consiste en deux étapes : l'étape d'encodage et l'étape de décodage. L'encodeur est utilisé pour mapper les données d'entrée dans une représentation cachée, et le décodeur est utilisé pour reconstruire les données d'entrée à partir de la représentation cachée.

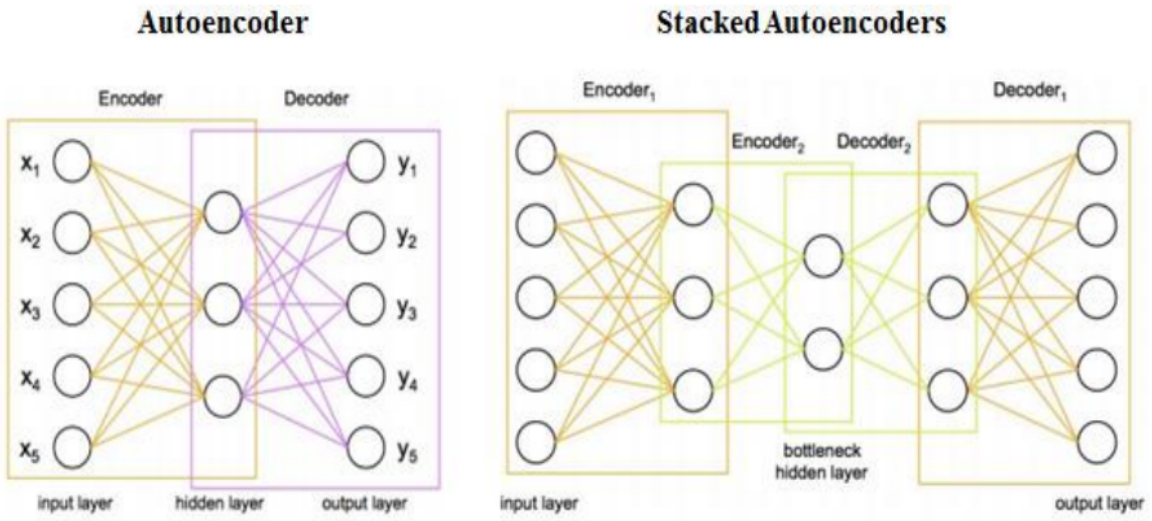


FIGURE 3.8 – Architecture des autoencodeurs [5]

Les SAEs sont généralement entraînés en deux étapes : la préformation et le réglage fin. Dans l'étape de préformation, chaque modèle d'autoencodeur est entraîné de manière non supervisée couche par couche, de bas en haut. Cette opération est répétée jusqu'à ce que les paramètres de toutes les couches cachées soient entraînés. Une fois que toutes les couches cachées sont entraînées, l'algorithme de rétropropagation est utilisé pour minimiser la fonction de coût et mettre à jour les poids avec un ensemble d'entraînement étiqueté afin d'effectuer un réglage fin.

3.3.4.4 Deep Belief Network (DBN)

D'après l'article de Bilal Jan et al. (2017) [28], le réseau de croyance profonde (Deep Belief Network, DBN) est un modèle qui peut apprendre la représentation des caractéristiques à partir de données étiquetées et non étiquetées en utilisant des techniques non supervisées et supervisées. L'architecture du DBN, comme illustré dans la Figure 3.9, comprend une couche d'entrée, des couches cachées et une couche de sortie, avec

deux couches connectées directement formant une machine de Boltzmann restreinte (RBM) [1].

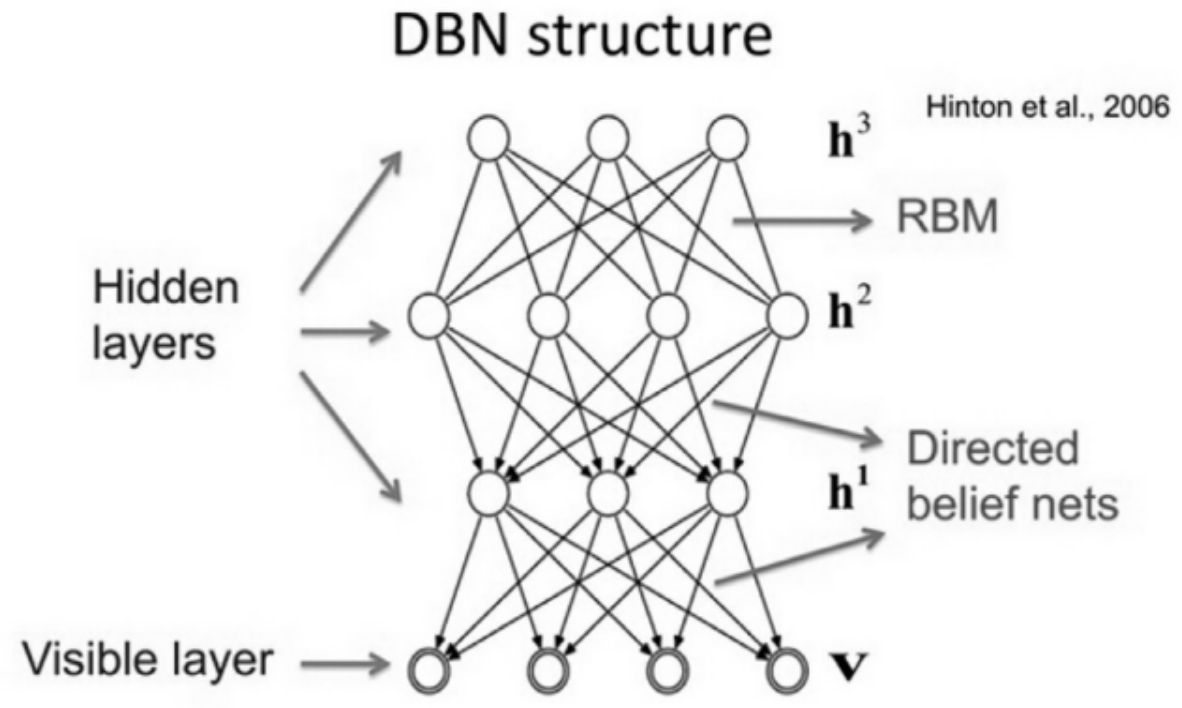


FIGURE 3.9 – Deep Belief Network [6].

Les chercheurs exploitent la puissance de la discrimination de DBN pour traiter efficacement les Big Data. Afin de gérer des quantités massives de données en un temps de traitement réduit, un modèle d'architecture basé sur les unités de traitement graphique (GPU) a été proposé. Cependant, l'implémentation de ce modèle pour les données à grande échelle présente divers défis, notamment le transfert de données entre la mémoire client et la mémoire globale. Une solution efficace consiste à stocker tous les paramètres et la solution d'entraînement dans la mémoire globale pendant la phase d'entraînement, tandis qu'un traitement parallèle des données permet des mises à jour simultanées sur chaque bloc d'informations. L'implémentation sur GPU démontre son efficacité lors de l'incorporation de plusieurs millions de paramètres dans RBM [28].

Le tableau suivant (**Tableau ??**) présente une comparaison entre quatre algorithmes de Deep Learning :

1. RBM comporte deux couches, les nœuds des deux couches étant entièrement connectés entre eux et aucune connectivité entre les nœuds au sein de la même couche

TABLE 3.4 – Comparaison entre les algorithmes Deep learning.

Références	Algorithme	Évolutivité	Efficacité	Précision	Confidentialité
DL-BCDSM [4]	CNN	oui	oui	oui	X
DL-BDA[28]	DBN	oui	X	oui	X
SDL-BDA [5]	SAEs	X	oui	X	X
PDRN-BDA)[29] DL-BDA[28]	RNN	oui	X	oui	X

En conclusion, le volume de données massives générées à une vitesse impressionnante dépasse les capacités des techniques d'apprentissage automatique traditionnelles pour fournir des solutions optimales aux ensembles de données complexes. Les techniques de deep learning ont réalisé des avancées significatives en abordant les problèmes de données en permettant l'identification de prédictions plus précises à partir de vastes quantités de données, en dévoilant des motifs et des structures complexes. Le deep learning utilise des réseaux de neurones artificiels pour effectuer un traitement multi-couches des données, permettant aux algorithmes d'extraire des caractéristiques de haut niveau à partir des données brutes. Cette approche simplifie non seulement le traitement des données, mais améliore également la précision et l'efficacité des modèles d'apprentissage automatique.

3.4 Comparaison entre Les algorithmes de chaque type du Marchine Learning

Une comparaison entre les différents types d'algorithmes de Machine Learning est présentée dans le Tableau ci-dessous :

TABLE 3.5 – Comparaison Entre Les algorithmes de chaque type du Machine Learning.

Type d'apprentissage	**Nom de l'algorithme**	**Année**	**Auteur(s)**	**Complexité**	**Domaines**
Supervised Learning	SVM trained multi layer neural network	2020	AlZubi	$O(N^3)$	Classification, Régression
Supervised Learning	Decision tree model	2018	Nair, et Shetty	O(n)	Classification, Régression
Supervised Learning	Regression algorithms with ensemble learning	2018	Asencio-cortés et al	$O(N^2)$	Régression
Unsupervised Learning	CLUBS-P	2020	Anni et al	$O(n\log(n))$	Clustering
Unsupervised Learning	Systèmes flous évolutifs multi-objectifs	2017	Pulgar-Rubio et al	O(n)	Clustering
Unsupervised Learning	Markov random fields	2018	IP et al	Non spécifier	Traitement et segmentation d'image
DL	RNN	2020	Hammou, Lahcen et Mouline	$O(N^2)$	Traitement du langage naturel
DL	DBN	2017	Bilal Jana et al	Non spécifier	Classification, Régression
DL	$CNN_{E}ndtoEnd$	2019	Li Shen et all et al	$O(N^2)$	Vision par ordinateur, Traitement d'image
SSL	Méthodes basées sur graphes	2005	Zhu	$O(N^3)$	Classification, Régression, Clustering
SSL	Méthode des k-means	1967	MacQueen	O(n)	Clustering
SSL	Self Organizing Map (SOM)	2003	Haykin	Non spécifier	Clustering, Réduction de dimensionnalité

3.5 Conclusion

Dans ce chapitre, nous avons exploré différentes approches de traitement du big data en utilisant des algorithmes de machine learning. Nous avons examiné les types d'algorithmes couramment utilisés, tels que l'apprentissage supervisé, non supervisé, semi-supervisé, ainsi que le deep learning.

L'analyse du big data nécessite des outils puissants pour extraire des informations significatives à partir de ces vastes ensembles de données. Le deep learning, en particulier l'utilisation des réseaux de neurones convolutifs (CNN), s'est révélé très prometteur dans ce domaine. Les CNN sont capables d'apprendre des caractéristiques complexes à partir de données brutes, ce qui permet une analyse approfondie et précise du big data.

Dans le chapitre suivant, nous proposons un algorithme de traitement du big data basé sur le deep learning et les CNN. Nous nous concentrons spécifiquement sur l'application de cet algorithme à la détection du cancer du sein. En nous inspirant des travaux de recherche existants, notamment l'article "Deep Learning to Improve Breast Cancer Detection on Screening Mammography" de Li Shen et al., nous développons notre propre méthode pour améliorer la détection précoce du cancer du sein à partir de grandes quantités de données.

Ce chapitre ouvre la voie à des possibilités passionnantes pour l'analyse du big data en utilisant le deep learning et les CNN. Les avancées continues dans ce domaine offrent des opportunités pour développer des modèles plus précis et efficaces, permettant ainsi d'extraire des informations précieuses à partir du big data.

Dans le chapitre suivant, nous détaillerons notre approche de traitement du big data utilisant le deep learning et les CNN pour la détection du cancer du sein.

Chapitre 4

Contribution & implémentation

4.1 Introduction

Ce chapitre constitue une exploration approfondie des différentes étapes du traitement du big data nécessaires pour la création de notre méthode de détection du cancer du sein. Notre attention se porte sur l'utilisation des algorithmes de deep learning, en particulier les principes du réseau de neurones à convolution (CNN) exposés dans l'article de référence précédent [4].

En nous inspirant des recherches précédentes et des principes du CNN, nous concevrons une méthode spécifique pour la détection du cancer du sein. Notre approche consistera en une analyse minutieuse des données volumineuses disponibles, telles que les images médicales, les données cliniques et d'autres sources pertinentes. Nous mettrons en œuvre des techniques avancées de prétraitement des données, d'extraction des caractéristiques et d'apprentissage profond pour améliorer la précision et la performance de notre système de détection.

Nous aborderons également les défis spécifiques liés à la manipulation du big data, tels que la gestion des données massives, les problèmes de stockage et les contraintes de calcul. Nous proposerons des solutions efficaces pour tirer pleinement parti de la puissance de traitement offerte par les infrastructures modernes, y compris l'utilisation de ressources informatiques parallèles et distribuées.

En conclusion, ce chapitre servira de guide exhaustif pour la mise en place des étapes nécessaires au traitement du big data dans le but de développer notre méthode de détection du cancer du sein. Grâce à l'utilisation du deep learning et des principes du CNN,

nous serons en mesure de réaliser une analyse approfondie des données massives, ce qui nous permettra d'améliorer significativement la précision et l'efficacité de la détection précoce du cancer du sein.

4.2 Objectifs de la recherche

Actuellement, la détection précoce du cancer du sein repose principalement sur l'expertise des radiologues et les programmes de dépistage par mammographie, mais ces méthodes présentent des limitations, notamment une interprétation humaine nécessaire et un taux élevé de faux positifs.

C'est pourquoi nous avons choisi de nous concentrer sur ce domaine pour appliquer nos méthodes de traitement des big data. Nous envisageons d'utiliser des techniques de deep learning (DL) pour développer un modèle capable d'automatiser et d'optimiser le processus d'évaluation des mammographies de dépistage. L'objectif est de fournir aux radiologues des outils d'aide à la décision basés sur l'intelligence artificielle (IA).

Les objectifs spécifiques du traitement des big data et des approches de DL dans le contexte de la classification du cancer du sein sont les suivants :

— **Réduire le nombre de faux positifs :** En réduisant les erreurs de diagnostic, nous pouvons atténuer l'anxiété inutile chez les patients et minimiser les procédures invasives telles que les biopsies. De plus, en rendant le dépistage plus accessible et moins dépendant de la disponibilité de radiologues hautement qualifiés, nous pouvons étendre les avantages de la détection précoce à une population plus large, notamment dans les régions où les ressources médicales sont limitées.

— **Améliorer la précision et l'efficacité de la classification du cancer du sein :** L'utilisation du traitement des big data et du DL vise à obtenir des modèles de classification du cancer du sein plus précis et plus efficaces. L'objectif est d'améliorer les performances de détection des lésions cancéreuses tout en réduisant les faux positifs et les faux négatifs.

— **Prédire le risque de cancer :** Utiliser le modèle pour évaluer le risque de cancer du sein en analysant les caractéristiques extraites des images mammographiques. Cela peut aider à identifier les femmes qui pourraient bénéficier de dépistages plus fréquents ou de stratégies de prévention spécifiques.

— **Améliorer les performances et réduire le temps de traitement :** Les techniques de DL permettent d'optimiser les performances de traitement des big data en utilisant des algorithmes parallèles et distribués, ce qui peut réduire considérablement le temps de traitement et améliorer l'efficacité globale du système.

— **Détecter des motifs complexes :** Les réseaux de neurones profonds, tels que les réseaux de neurones à convolution (CNN), sont capables de détecter des motifs complexes tels que des contours, des textures ou des formes dans différentes parties de l'image. L'utilisation de couches profondes permet de découvrir des motifs plus abstraits et complexes, rendant ainsi l'analyse des images plus précise et efficace.

En poursuivant ces objectifs, nous visons à améliorer significativement la détection précoce du cancer du sein en exploitant les possibilités offertes par le traitement des big data et les approches de DL, en particulier les réseaux de neurones à convolution.

4.3 Cahier de charge

Le cancer du sein est la forme de cancer la plus répandue et la principale cause de décès chez les femmes à l'échelle mondiale. Chaque année, environ 2 millions de femmes sont diagnostiquées avec un cancer du sein, ce qui en fait une maladie très répandue [30]. Selon l'OMS, en 2020, on estime qu'il y a eu 685 000 décès liés au cancer du sein dans le monde. En Algérie, environ 14 000 nouveaux cas de cancer du sein sont diagnostiqués chaque année, soulignant l'importance et la prévalence de cette maladie dans le pays [1] En 2020, le cancer du sein constituait 11,7 % de l'ensemble des nouveaux cas de cancer signalés à l'échelle mondiale [2]

La Figure 4.1 illustre la taxonomie de ce cahier des charges, établie suite à des échanges avec des professionnels de la santé, notamment des médecins et des étudiants en médecine de l'université Ferhat Abbas Sétif. Cette étape nous permet de développer une compréhension approfondie de notre domaine d'étude et de proposer les solutions les plus adaptées à notre problématique.

1. https ://www.allodocteurs.africa/le-cancer-du-sein-gagne-du-terrain-en-algerie-6671.html
2. https ://news.un.org/fr/story/2021/02/1088502

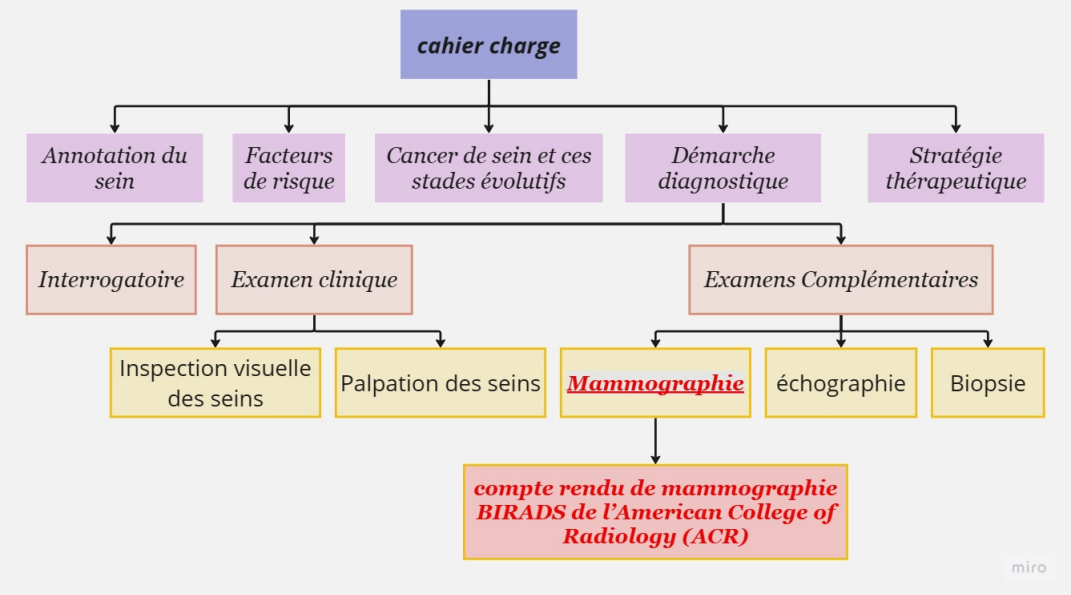


FIGURE 4.1 – Organigramme du cahier de charge.

4.3.1 Annotation du sein

La fonction principale du sein est de produire du lait pour nourrir un nouveau-né. Il se compose d'une glande mammaire, de fibres de soutien et de tissu adipeux. La glande mammaire est divisée en 15 à 20 sections appelées lobules, qui sont reliés à des canaux aboutissant au mamelon (voir Figure 4.2).

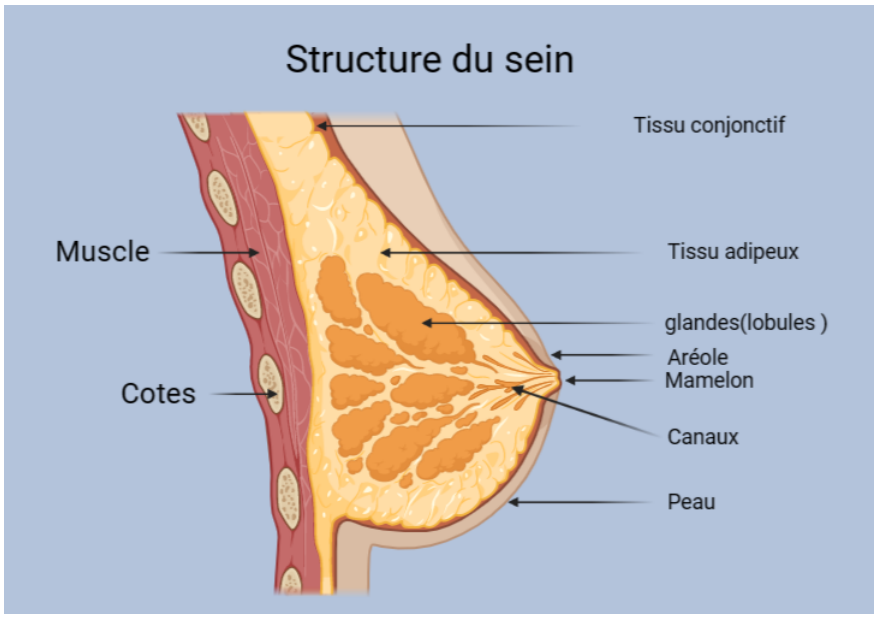


FIGURE 4.2 – Structure du sein.

Les seins peuvent être sujets à différentes conditions médicales, telles que le cancer du sein, les kystes mammaires, les infections, les douleurs et les changements hormonaux.

Il est donc essentiel de surveiller régulièrement la santé des seins et de consulter un médecin en cas de changement ou d'anomalie dans cette zone.

4.3.2 Cancer du sein et ses stades évolutifs

Le cancer du sein est une tumeur maligne de la glande mammaire, caractérisée par une multiplication incontrôlée des cellules. Les cancers du sein sont classés en plusieurs stades évolutifs en fonction de leur taille, de leur propagation dans le tissu mammaire et de leur extension dans les tissus environnants. Voici les principaux stades :

— **Cancers in situ :** Les cancers in situ se caractérisent par la présence de cellules cancéreuses à l'intérieur des canaux galactophores ou des lobules, sans que la tumeur ait franchi la membrane basale qui les entoure ni infiltré le tissu voisin. On distingue le carcinome canalaire in situ et le carcinome lobulaire in situ. Environ 85 à 90% des cancers in situ sont des carcinomes canalaires in situ. La maladie de Paget du mamelon est également un type de carcinome canalaire in situ qui peut se propager à l'aréole ou au tissu mammaire plus profond.

— **Adénocarcinomes infiltrants :** Lorsque les cellules cancéreuses ont infiltré le tissu mammaire environnant, on parle de cancer infiltrant ou de carcinome infiltrant. Les cancers infiltrants sont plus agressifs que les cancers in situ et peuvent se propager aux tissus environnants, notamment aux ganglions lymphatiques. Les carcinomes canalaires infiltrants représentent environ 80% des cas, tandis que les carcinomes lobulaires infiltrants, plus rares, représentent environ 10

— **Cancers localement avancés :** Les cancers localement avancés se caractérisent par une propagation importante de la tumeur dans les tissus environnants tels que la peau, les muscles ou les ganglions lymphatiques. Ils sont généralement plus difficiles à traiter que les cancers à un stade précoce.

— **Cancers métastatiques :** Les cancers métastatiques se réfèrent aux cancers qui se sont propagés à d'autres parties du corps. Ils sont considérés comme les stades les plus avancés du cancer du sein et nécessitent souvent un traitement palliatif visant à soulager les symptômes et à améliorer la qualité de vie. La Figure 4.3 illustre le stade IV, caractérisé par une hétérogénéité cellulaire et une propagation des cellules malignes au-delà de l'emplacement initial de la tumeur. À ce stade, les cellules cancéreuses métastasent vers les ganglions lymphatiques axillaires et d'autres organes

distants tels que le cerveau, le foie, les os et les poumons.

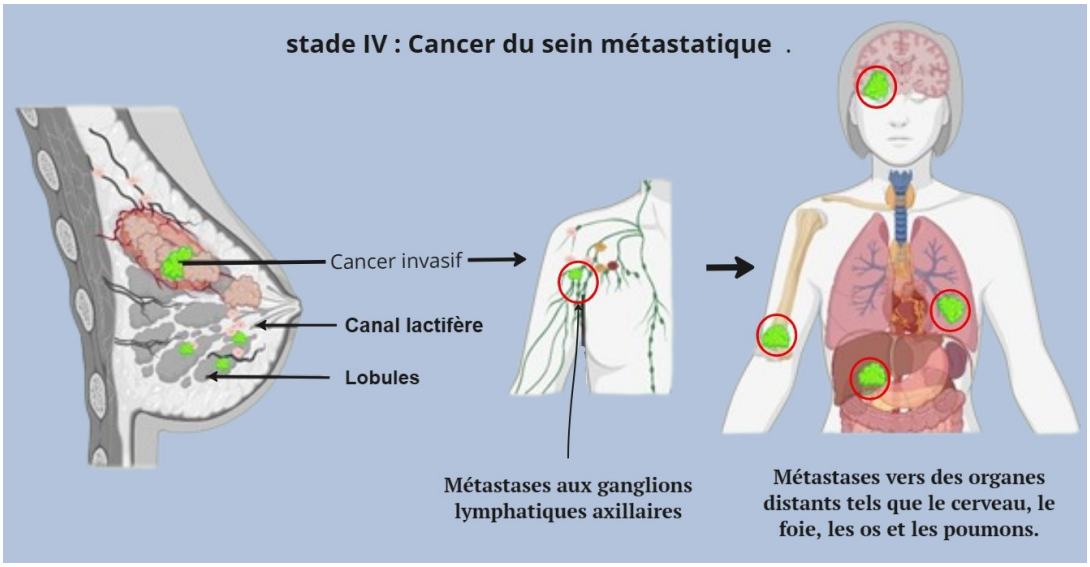


FIGURE 4.3 – Cancers métastatiques.

4.3.3 Facteurs de risque

Plusieurs facteurs de risque importants sont associés au cancer du sein :

— **Âge :** Plus de 78% des cas surviennent chez les femmes de plus de 50 ans[30].

— **Sexe féminin :** Moins de 1% des cas sont observés chez les hommes[30].

— **Prédisposition génétique :** Les mutations des gènes BRCA1 ou BRCA2 représentent environ 5% à 10% des cas de cancer du sein.

— **Antécédents personnels :** Les antécédents de cancer du sein invasif, de carcinome canalaire in situ, d'hyperplasie lobulaire atypique ou de cancer lobulaire in situ augmentent le risque [31].

— **Exposition à l'irradiation thoracique :** Une exposition antérieure à des doses élevées d'irradiation thoracique, notamment pour le traitement de la maladie de Hodgkin, est un facteur de risque [31].

— **Traitement hormonal substitutif et ménopause :** L'utilisation de traitements hormonaux substitutifs et la ménopause peuvent accroître le risque.

— **Nulliparité ou faible nombre d'enfants :** Le fait de n'avoir jamais eu d'enfant ou d'avoir eu peu d'enfants est associé à un risque plus élevé.

4.3.4 Démarche diagnostique

La démarche diagnostique englobe les différentes étapes permettant d'identifier la maladie ou le trouble dont souffre un patient. Bien que ces étapes puissent varier en fonction de la pathologie et des symptômes présentés, certaines d'entre elles sont généralement communes à la plupart des démarches diagnostiques.

Il existe plusieurs situations où la présence d'un cancer du sein chez une femme peut être suspectée, notamment :

— La détection d'anomalies lors d'un examen clinique, telles que la présence d'un nodule palpable, d'une sécrétion unipore mamelonnaire sérosanglante ou de la maladie de Paget du mamelon.

— Le dépistage mammographique pour les femmes âgées de 50 à 74 ans.

Quelle que soit la méthode de détection, une évaluation diagnostique est toujours nécessaire. Cela comprend un examen clinique, des examens radiologiques et un diagnostic histologique pour confirmer ou infirmer la présence d'un cancer du sein.

4.3.4.1 Interrogatoire

L'interrogatoire consiste à poser des questions sur les antécédents médicaux, les symptômes et les facteurs de risque du patient pour le cancer du sein. Cette étape permet de mieux comprendre la situation du patient et d'orienter la démarche diagnostique vers les examens complémentaires nécessaires.

4.3.4.2 Examen clinique

L'examen clinique revêt une importance particulière dans la démarche diagnostique du cancer du sein. Il comprend une inspection visuelle et une palpation des seins et des aisselles afin de détecter d'éventuelles anomalies.

1. **Inspection visuelle des seins :** Cette étape consiste à examiner visuellement les seins à la recherche de signes cliniques tels qu'une tuméfaction, un nodule, des changements de taille ou d'aspect, une asymétrie, une voussure ou une rétraction d'un sein, ainsi que la présence de ganglions dans la région axillaire et sus-claviculaire.
2. **Palpation des seins :** La palpation des seins permet une exploration minutieuse des quadrants externes et internes, ainsi que de la région sous et rétro-aréolaire. Cette

étape vise à détecter d'éventuels nodules malins associés au cancer du sein. Des techniques de palpation sont utilisées pour rechercher toute anomalie, telle qu'une tuméfaction indolore, de consistance dure, aux contours réguliers ou irréguliers, ainsi que des signes tels qu'une rétraction ou une distorsion du mamelon, un pli cutané, un épaississement de l'aréole avec une peau d'orange ou une inflammation cutanée.

3. **Palpation des aires ganglionnaires :** La palpation des aires ganglionnaires, notamment dans la région axillaire et sus-claviculaire, permet de rechercher la présence de ganglions suspects qui pourraient indiquer une extension du cancer du sein aux ganglions lymphatiques.

4.3.4.3 Examens complémentaires

En cas de suspicion après l'examen clinique, des examens complémentaires sont prescrits pour confirmer ou exclure la présence d'un cancer du sein. Les examens les plus couramment utilisés sont la mammographie et l'échographie mammaire.

1. **Mammographie** La mammographie est une technique d'imagerie médicale qui utilise des **rayons X** pour produire des images des seins. C'est l'examen de référence pour le dépistage et le diagnostic du cancer du sein. Pendant l'examen, le sein est placé entre deux plaques et une faible dose de rayonnement est utilisée pour prendre des images sous différents angles (**voir figure ??**). Les images sont ensuite examinées par un radiologue afin de détecter d'éventuelles anomalies ou signes de cancer. La mammographie est recommandée tous les deux ans pour les femmes âgées de 50 à 74 ans dans le cadre d'un dépistage régulier. Pour les femmes présentant des symptômes tels que des masses ou des douleurs mammaires, une mammographie peut être réalisée à tout âge. Bien que la mammographie soit un examen essentiel pour la détection précoce du cancer du sein, il peut y avoir des limites et des résultats faussement positifs. Il est important de discuter avec son médecin de la nécessité et de la fréquence de la mammographie en fonction de ses antécédents médicaux et de ses facteurs de risque.

   Compte rendu de mammographie

   Le compte rendu de mammographie doit fournir des informations sur les calcifications, y compris leur nombre, leur taille, leur localisation, leur forme, leur distribu-

tion spatiale, leur évolution dans le temps et la présence de signes associés.

Le système BIRADS de l'American College of Radiology (ACR) est utilisé pour classer le degré de suspicion de malignité (**voir Tableau ??**). Les catégories ACR vont de 0 à 5 et permettent de guider les actions nécessaires en fonction des résultats de la mammographie, telles que des investigations complémentaires ou une biopsie.

TABLE 4.1 – Classification ACR pour le dépistage du cancer du sein

Catégorie	**Image**	**Risque de cancer**	**Action**
ACR 0	L'évaluation mammographique est incompléte	N/A	Des investigations complémentaires sont nécessaires
ACR 1	Mammographie normale	0%	Retour au dépistage
ACR 2	Constatations bénignes	0%	Retour au dépistage
ACR 3	Anomalie probablement bénigne	>2% de risque de malignité	Proposition d'une surveillance initiale à court terme (de 4 à 6 mois)
ACR4	**Anomalie suspecte :**		Une biopsie doit être envisagée
ACR 4A	-Valeur Prédictive Positive Faible	2-10%	
ACR 4B	-Valeur Prédictive Positive Intermédiaire	10-50%	
ACR 4C	-Valeur Prédictive Positive Forte	50-90%	
ACR 5	Anomalie évocatrice d'un cancer (haute probabilité de malignité)	$\geq$95%	Biopsie

2. **Échographie**

Lorsqu'un cancer du sein est suspecté, l'échographie est un examen complémentaire important pour caractériser une lésion identifiée lors de la mammographie. Cet examen permet d'obtenir une image plus précise de la lésion, de sa taille, de sa forme et de sa structure. L'échographie est également utile chez les femmes enceintes ou les femmes jeunes ayant des seins denses, pour lesquelles la mammographie peut être moins efficace. L'examen est indolore, rapide et ne nécessite pas d'irradiation.

Lors d'une échographie, les signes évocateurs de malignité comprennent une image hypoéchogène (plus sombre que les tissus environnants), une forme irrégulière avec des contours flous, une taille inférieure à une masse palpable, une ombre postérieure et une augmentation du flux sanguin (détectable par Doppler).

Ces signes ne sont pas spécifiques au cancer et peuvent également être observés dans des lésions bénignes, mais ils doivent être pris en compte pour orienter le diagnostic et décider d'une investigation complémentaire, telle qu'une biopsie.

4.3.4.4 Examen histologique : importance de la biopsie

L'examen histologique joue un rôle crucial dans la confirmation du diagnostic du cancer du sein. Il est réalisé à travers une biopsie ou une ponction cytologique, en fonction de la nature et de la localisation de la lésion.

— Si la lésion est palpable, la biopsie ou la ponction cytologique peut être effectuée en utilisant la détection tactile pour guider l'aiguille ou l'instrument.

— Dans le cas d'une lésion non palpable, la biopsie ou la ponction est guidée par échographie ou mammographie.

4.3.5 Approche thérapeutique

Le traitement du cancer du sein comprend plusieurs options, telles que la chirurgie, la radiothérapie, la chimiothérapie et l'hormonothérapie. Les décisions concernant les modalités thérapeutiques appropriées et leur séquence d'administration sont discutées lors d'une réunion de concertation pluridisciplinaire.

4.4 Ensembles de données (Dateset)

Récemment, les systèmes d'aide à la détection et au diagnostic du cancer du sein assistés par ordinateur (CAD) sont de plus en plus utilisés. Les ensembles de données d'images mammographiques jouent un rôle crucial dans le développement de versions améliorées de ces systèmes CAD[32].

Dans leur livre, **Khalid Shaikh et al. (2021)[32]**, présentent dans la section 6 plusieurs ensembles de données mammographiques populaires disponibles dans la littérature, tels que Nbreast, BancoWeb LAPIMO, DDSM, etc. Ces ensembles de données, ainsi que

d'autres ensembles de données sur le cancer du sein, sont utilisés dans de nombreux projets de recherche et sont étudiés dans une grande partie de la littérature.

Après des recherches approfondies pour choisir un ensemble de données adapté à nos besoins, nous avons trouvé plusieurs ensembles de données publics disponibles sur Kaggle et d'autres plateformes. Nous avons choisi l'ensemble de données du concours **RSNA Breast Cancer Detection Challenge (2023)** organisé par la Radiological Society of North America (RSNA) pour la détection du cancer du sein à partir de mammographies de dépistage.

Cet ensemble de données a été fourni par des programmes de dépistage par mammographie en Australie et aux États-Unis. Il comprend des étiquettes détaillées avec les évaluations des radiologistes et les résultats de la pathologie de suivi pour les tumeurs malignes suspectées.

Voici une description des fichiers inclus dans l'ensemble de données :

4.4.1 Ensemble de données RSNA Breast Cancer Detection

Les mammographies disponibles dans l'ensemble de données sont au format DICOM. On peut s'attendre à environ 8 000 patients dans l'ensemble de test caché, avec généralement mais pas toujours 4 images par patient. Il est important de noter que de nombreuses images utilisent le format JPEG 2000, ce qui peut nécessiter le chargement de bibliothèques spéciales.

4.4.2 [train/test].csv

Le fichier [train/test].csv contient les métadonnées pour chaque patient et chaque image. Le tableau suivant présente les colonnes présentes dans le fichier CSV. Nous avons ajouté la colonne "donnée" pour spécifier si les données sont utilisées pour l'entraînement ou la validation du modèle.

TABLE 4.2 – Récapitulatif des attributs de fichier (**[train/test].csv**) et de leur disponibilité pour l'entraînement ou le test.

Attribut	**Description**	**Données (Entraînement /Test)**
site_id	Code ID pour l'hôpital source	Entraînement et Test
patient_id	Code ID pour le patient	Entraînement et Test
image_id	Code ID pour l'image	Entraînement et Test
laterality	Indique si l'image est du sein gauche ou droit	Entraînement et Test
view	L'orientation de l'image	Entraînement et Test
age	L'âge du patient en années	Entraînement et Test
implant	Indique si le patient a des implants mammaires	Entraînement
density	Une évaluation de la densité du tissu mammaire	Entraînement
machine_id	Un code ID pour l'appareil d'imagerie	Entraînement et Test
cancer	Indique si le sein est positif pour un cancer malin	Entraînement
biopsy	Indique si une biopsie de suivi a été réalisée sur le sein	Entraînement
invasive	Si le sein est positif pour un cancer, indique si le cancer s'est révélé invasif	Entraînement
BIRADS	Évaluation du sein : 0 si suivi nécessaire, 1 si négatif pour le cancer, 2 si normal	Entraînement
prediction_id	L'ID correspondant à la ligne de soumission associée	Test
difficult_ negative_ case	Vrai si le cas était exceptionnellement difficile	Entraînement

4.5 Pré-traitement des données

Le pré-traitement des données comporte deux étapes cruciales pour assurer la qualité des données : le nettoyage des images et la préparation du fichier CSV. Ces étapes visent à préparer les données pour une utilisation efficace dans les modèles d'apprentissage automatique.

4.5.1 Prétraitement des images

1. **Sélection d'un sous-ensemble de données** Dans un premier temps, nous avons dû faire une sélection restreinte des données du Data Set. Cette décision a été motivée par nos contraintes de ressources, car le volume total des données atteignait 314 Go.

   Par conséquent, nous avons choisi de travailler avec un sous-ensemble de données réduit, converti au format PNG, afin de pouvoir utiliser TensorFlow, qui ne peut pas lire directement les fichiers DICOM.

2. **Nettoyage des données**

   Nous avons procédé à une vérification des images pour identifier les éventuelles corruptions ou absences de données, et les avons supprimées du Data Set afin d'éviter tout problème lors de l'apprentissage automatique.

   La **Figure 4.4** illustre deux exemples issus de notre Data Set. Nous pouvons observer la différence entre ces deux images : **l'image A** de qualité médiocre a été supprimée pour éviter toute perturbation lors de l'apprentissage automatique, tandis que **l'image B** sera utilisée pour l'entraînement de nos modèles.

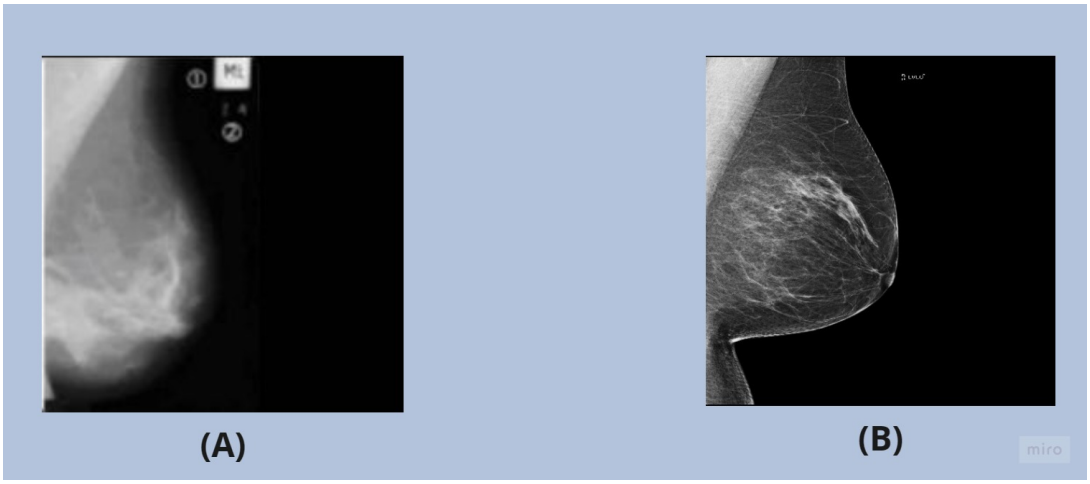


FIGURE 4.4 – Nettoyage des images du Data Set

3. **Normalisation**

   La normalisation en big data consiste à organiser les données volumineuses et complexes de manière structurée et standardisée. L'objectif principal de la normalisation est d'éliminer les redondances, de réduire la taille des données et d'améliorer l'efficacité des opérations de traitement.

4. **Redimensionnement**

   Toutes les images ont été redimensionnées pour avoir une taille fixe de **512*512 pixels**, afin de les rendre compatibles avec notre modèle qui nécessitait une taille d'entrée constante.

4.5.2 Prétraitement du fichier CSV

1. **Chargement des annotations** Les fichiers CSV contiennent souvent diverses informations associées aux mammographies, telles que leur classification (anormale / normale), leur emplacement dans le corps et leur taille.

2. **Conversion en format numérique**

   Les annotations sont généralement fournies sous forme textuelle, mais il est souvent nécessaire de les convertir en format numérique (par exemple, 1 pour une anomalie présente et 0 pour une anomalie absente) afin de pouvoir les utiliser directement dans notre réseau neuronal.

3. **Fusion avec les images**

   Enfin, nous avons fusionné les annotations avec notre jeu d'images, de sorte que chaque ligne du fichier CSV corresponde à l'image correspondante dans le Data Set.

Ces étapes de nettoyage et de préparation des données sont essentielles pour garantir la qualité et la pertinence des informations utilisées dans notre modèle d'apprentissage automatique. Elles permettent de minimiser les biais, d'éliminer les erreurs et de préparer les données de manière optimale pour l'entraînement et la validation du modèle.

4.6 Stockage et gestion des données

Dans le cadre de la gestion et de l'analyse des données, nous avons été confrontés à d'importantes limitations en termes de stockage et de capacité de traitement. Malgré nos efforts pour utiliser des outils avancés tels que Hadoop et Spark, il est devenu évident que nos machines ne disposaient pas des ressources nécessaires pour effectuer des analyses de données efficaces. Face à ce défi, nous avons exploré différentes solutions et opté pour une approche basée sur le cloud.

Nous avons choisi d'utiliser Google Drive comme plateforme de stockage dédiée et avons établi une connexion solide entre Kaggle et Colab. Cette configuration nous permet d'utiliser directement le code provenant de Kaggle dans Colab, tout en conservant nos ensembles de données sur Google Drive.

Cette solution astucieuse a considérablement amélioré notre capacité à gérer et à analyser de vastes ensembles de données, en nous offrant une flexibilité accrue et en contournant les limitations matérielles de nos machines. Grâce à cette approche, nous pouvons tirer pleinement parti des fonctionnalités offertes par le cloud, notamment en termes de stockage, de puissance de calcul et de collaboration.

4.7 Exploration des données

L'exploration des données revêt une importance primordiale dans l'analyse de données, car elle permet d'examiner les caractéristiques et la distribution de l'ensemble de données. Cette étape joue un rôle crucial dans l'identification des relations entre les différentes variables, la détection des valeurs aberrantes ou manquantes, ainsi que la détermination des techniques d'analyse appropriées pour un ensemble de données spécifique.

Dans cette section, nous avons procédé comme suit :

1. Nous avons utilisé la bibliothèque Python Pandas pour charger les données et créer un tableau de données (DataFrame) contenant toutes les informations nécessaires. Le tableau suivant (**Tableau 1**) présente la répartition des images selon différentes catégories :
2. Nous avons étudié la répartition des classes malignes et bénignes en utilisant diffé-

TABLE 4.3 – Répartition des images selon les catégories

Category	Nombre Image
Implant = 1	1,477
Cancer = 0	53,548
Cancer = 1	1,158
Biopsy = 1	2,969
Invasive = 1	818
Cancer = 1 & Invasive = 1	818
Biopsy = 1 & Cancer = 0	1,811
Biopsy = 1 & Cancer = 1	1,158
Total Images	54,706

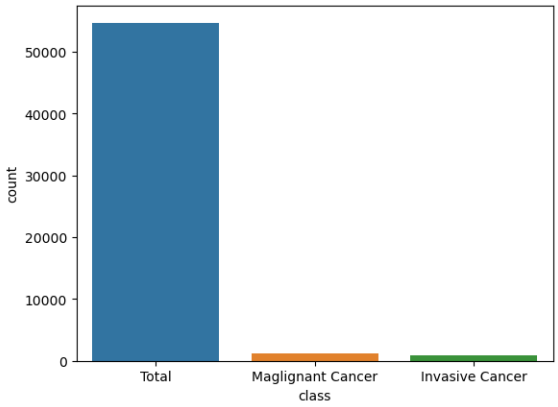
rents types de graphiques afin d'avoir une vision globale de chaque classe.

Nous avons analysé la relation entre certaines variables, comme la taille des images mammographiques, et leur nature maligne ou bénigne en utilisant des diagrammes de dispersion 2D avec une coloration différente pour chaque classe (**Figure 4.5**).

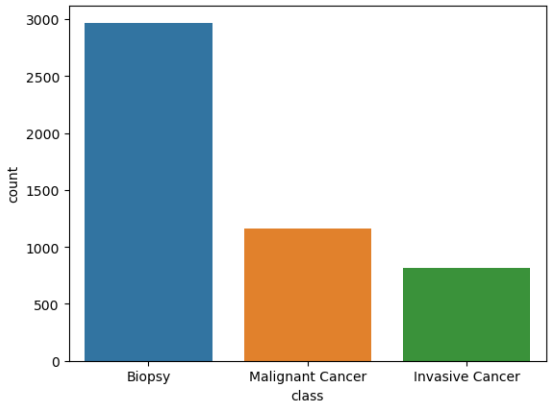
Le Tableau suivant représente Description de la **Figure 4.5**

TABLE 4.4 – Description de la **Figure 4.5**

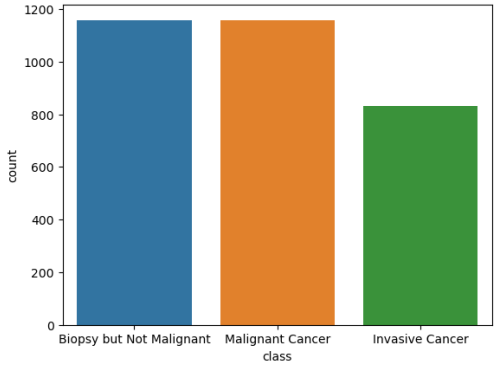
graphe	Bleu	Orange	Vers
(a)	**Totale Image**	**Maglignant Cancer** 'cancer'=1	**'Invasive Cancer'** 'cancer' = 1 et 'invasive'= 1
(b)	**Biopsy** 'biopsy'= 1	**'Malignant Cancer** 'cancer'= 1	**Invasive Cancer** 'cancer' = 1 et 'invasive' = 1
(c)	**Biopsy but Not Malignant** 'biopsy'= 1 et 'cancer'= 0	**'Malignant Cancer'** 'cancer' = 1	**'Invasive Cancer'** 'cancer' = 1 et 'invasive' = 1
(d)	**'Biopsy but Not Malignant'** 'biopsy'= 1 et 'cancer = 0	**' Malignant Cancer'** 'cancer'= 1	**'Invasive Cancer'** 'cancer'= 1 et 'invasive' = 1



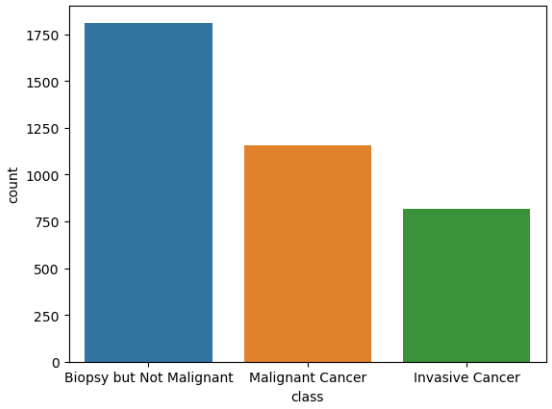
(a) La plupart des cas sont des cancers normaux ou non malins. Ainsi, les médecins négligent parfois le cancer.



(b) Environ 3000 patients ont subi une biopsie et un cancer malin a été découvert chez certains d'entre eux



(c) Généralement, le cancer invasif est confirmé par biopsie, et non par mammographie. Peut-être est-il également extrêmement difficile pour l'IA de détecter un cancer invasif à partir d'une mammographie



(d) 60 % des biopsies ont abouti à un cancer non malin.

FIGURE 4.5 – Présentation graphique des caractéristiques distinctives de l'ensemble de données sur le cancer du sein

3. Nous avons effectué une augmentation artificielle des données (data augmentation) pour équilibrer le nombre d'échantillons par classe, ce qui améliore la qualité de l'apprentissage.

   En résumé, cette exploration des données permet aux analystes de mieux comprendre les différentes propriétés statistiques essentielles du Data Set avant même l'utilisation effective des algorithmes d'analyse. Cela aide à identifier tout déséquilibre potentiel entre les classes malignes et bénignes, ainsi qu'à repérer toute valeur aberrante ou manquante qui pourrait avoir un impact négatif sur l'apprentissage du modèle.

4.8 Traitement des données d'images

Le traitement des données d'images est une étape cruciale dans le domaine de l'apprentissage automatique et de la vision par ordinateur. Il comprend plusieurs opérations visant à préparer les images pour l'entraînement et la validation des modèles. Dans cette section, nous avons divisé cette étape en deux parties : la division des jeux de données et la création d'un générateur.

4.8.1 Division des jeux de données

Notre code comprend deux parties qui concernent la division des données en ensembles d'entraînement et de validation, en veillant à ce que les images normales et les images de cancer soient réparties de manière équilibrée. Cette division est essentielle pour évaluer les performances du modèle sur des données non vues lors de l'entraînement et pour éviter le surapprentissage.

```
train_df, val_df = train_test_split(DF_train, test_size=0.30,
    random_state=2018, stratify=DF_train[['cancer']])

print('train', train_df.shape[0], 'v alidation', val_df.shape[0])
print('train', train_df['cancer'].value_counts())
print('validation', val_df['cancer'].value_counts())
train_df.sample(1)
```

Dans notre code, nous avons utilisé la fonction "train_test_split" pour diviser les données en deux parties : un ensemble d'entraînement (70%) et un ensemble de validation (30%). Pour garantir une répartition similaire des images malignes et bénignes, nous avons stratifié la distribution des classes dans les deux ensembles. Ensuite, nous avons sélectionné les images normales à partir des données d'entraînement et les images de cancer à partir des données de validation.

Les résultats de l'exécution du code nous fournissent des informations détaillées sur la répartition des données dans les ensembles d'entraînement et de validation. Le tableau suivant (**Tableau 4.5**) présente ces informations :

TABLE 4.5 – Division Datasets

Datasets	Cancer	Normal	Total (échantillons)
Entraînement (Train)	811	810	1621
Validation (test)	347	348	695
Total (échantillons)	1158	1158	2316

Les chiffres obtenus montrent que les images normales et les images de cancer sont réparties de manière équitable entre les ensembles d'entraînement et de validation, avec des proportions similaires pour les deux classes.

Dans la deuxième partie du traitement des données d'images, nous avons créé des générateurs de données pour faciliter l'entraînement de notre modèle TensorFlow. Ces générateurs ont été utilisés pour préparer les ensembles d'entraînement et de validation, qui contiennent à la fois des images normales et des images de cancer du sein. L'objectif est de rendre les données prêtes à être utilisées dans l'apprentissage automatique.

```
train_datagen = ImageDataGenerator(rescale = 1./255.,zoom_range =
    0.2)
val_datagen = ImageDataGenerator(rescale = 1./255.,)
```

```
 train_path = 'dataset/train'
val_path = 'dataset/val'

train_generator = train_datagen.flow_from_directory(
    train_path,
    target_size = (512, 512),
    batch_size = 32,
    class_mode = 'binary'
)
validation_generator = val_datagen.flow_from_directory(
        val_path,
        target_size = (512, 512),
        batch_size = 16,
        class_mode = 'binary'
)
```

Par la suite, des générateurs de données d'images ont été instanciés en utilisant les classes ImageDataGenerator fournies par TensorFlow Keras. Ces générateurs offrent

une solution pratique pour charger un grand nombre d'images à partir du disque dur sans avoir à les charger en mémoire vive en une seule fois. De plus, les générateurs appliquent une augmentation aléatoire sur chaque image lors de leur chargement, ce qui permet au modèle d'apprendre à généraliser ses prédictions de manière plus efficace.
Avant d'être utilisées par le modèle de Deep Learning, les images sont redimensionnées à une taille standard de 512x512 pixels. De plus, la classe "class_mode='binary" est spécifiée, ce qui indique qu'il y a seulement deux classes possibles : "normal" ou "cancer". Par conséquent, notre modèle est configuré avec une couche finale utilisant la fonction d'activation sigmoid, qui retourne une valeur de 0 ou 1 pour chaque image.
De même, le générateur "validation_generator" est créé avec les mêmes paramètres, mais avec une taille de lot (batch size) de 16 pour la phase de validation. Ces générateurs de données nous permettent de charger les images de manière dynamique pendant l'entraînement du modèle, ce qui facilite la gestion d'ensembles de données volumineux et leur utilisation efficace.

Résultat :

Les sorties des générateurs fournissent des informations sur le nombre total d'images présentes dans chaque ensemble et le nombre de classes détectées. Dans notre cas, le générateur d'entraînement a trouvé 1224 images appartenant à 2 classes, tandis que le générateur de validation a trouvé 606 images appartenant à 2 classes. Veuillez noter que le nombre de classes peut varier en fonction de votre propre ensemble de données.

En conclusion :

Pour traiter les données d'images, nous avons d'abord divisé notre ensemble de données en ensembles distincts : l'ensemble d'entraînement et l'ensemble de validation. Ensuite, nous avons utilisé la classe ImageDataGenerator de TensorFlow pour créer des générateurs de données d'images. Ces générateurs nous permettent de charger et de prétraiter dynamiquement les images pendant l'apprentissage du modèle, en appliquant des transformations telles que la mise à l'échelle des pixels et l'augmentation des données. Cela facilite le processus d'entraînement en permettant une gestion efficace des ensembles de données volumineux.

4.9 Choix des algorithmes et des techniques :

Dans cette section, nous avons sélectionné l'algorithme du réseau de neurones convolutifs (CNN) pour l'analyse de nos données, car il est adapté à nos objectifs.

4.9.1 Généralités sur les réseaux de neurones convolutifs

La partie concernant les réseaux de neurones convolutifs (CNN) est divisée en deux sous-parties : la conception des CNN et leurs diverses architectures.

4.9.1.1 Conception des CNN

Un réseau de neurones convolutif (CNN), également connu sous le nom de réseau de convolution, est un type d'algorithme de deep learning largement utilisé pour l'analyse d'images, notamment dans des tâches telles que la classification et la segmentation. Il est composé de plusieurs couches, dont au moins une est une couche de convolution. La couche de convolution utilise des filtres spécifiques pour détecter des motifs dans l'image et préserver les caractéristiques spatiales.

Lors de l'apprentissage des réseaux de neurones convolutifs (CNN), la taille spatiale du volume d'entrée est généralement réduite en appliquant des fonctions d'activation aux résultats des couches de convolution, ou en combinant les résultats des couches de convolution avec des couches de pooling. Cette réduction progressive de la taille spatiale permet d'augmenter l'abstraction des caractéristiques et de mieux représenter les informations essentielles dans les couches suivantes du réseau. Finalement, le réseau atteint une couche entièrement connectée où la classification ou la segmentation finale est réalisée.

Voici une reformulation du texte en conservant la liste LaTeX et la figure :

Il existe plusieurs composants clés dans les réseaux de neurones convolutifs (CNN) :

1. **Optimisation** : Les optimiseurs sont des algorithmes utilisés pour ajuster les paramètres d'un modèle d'apprentissage automatique afin de minimiser une fonction de coût. En minimisant les erreurs, on améliore la précision de la classification. Parmi les optimiseurs couramment utilisés, on trouve le gradient stochastique descendant (SGD), Adam, Adagrad et RMSProp, entre autres.
2. **Max et average pooling (regroupement maximum et moyen)** : Pour résoudre le

problème de l'entraînement lent et coûteux en termes de calcul des grands canaux de **feature map** avec des modèles de convolution de base, les chercheurs appliquent le max-pooling et l'average pooling à la sortie des couches de convolution. Le max-pooling sélectionne la valeur maximale des pixels dans une fenêtre 2D (voir Figure 4.6), tandis que l'average pooling prend la moyenne des valeurs de tous les pixels dans une fenêtre 2D. Ces opérations de pooling fournissent une invariance aux translations mineures dans les données d'entrée et réduisent la taille des **feature map**. De plus, ces opérations réduisent le nombre de paramètres entraînables pour atténuer le surajustement et réduire le temps de calcul.

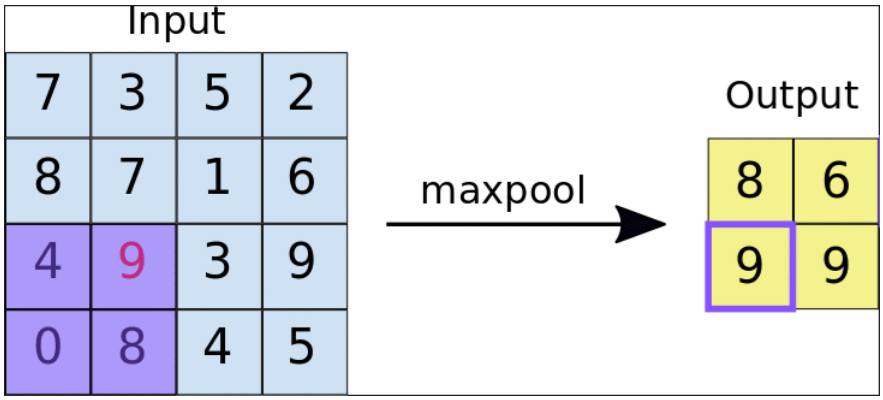


FIGURE 4.6 – Max-pooling

3. **Fonctions d'activation** : Après l'opération de convolution, une fonction d'activation non linéaire est généralement appliquée à chaque carte d'activation pour introduire de la non-linéarité dans le modèle. Des exemples de fonctions d'activation couramment utilisées sont ReLU (Rectified Linear Unit) et sigmoid. Ces fonctions permettent d'introduire des non-linéarités dans le modèle et d'améliorer sa capacité à capturer des relations complexes entre les caractéristiques.
4. **Dropout** : Le dropout est une technique de régularisation utilisée pour éviter le surapprentissage, en particulier dans les réseaux neuronaux. L'idée derrière le dropout est de supprimer aléatoirement certains nœuds à chaque itération en les mettant à zéro. Ainsi, le modèle ne peut pas trop dépendre de nœuds spécifiques, ce qui le rend plus généralisable. Pendant l'entraînement, le dropout sélectionne de manière aléatoire un sous-ensemble de nœuds dans chaque couche et met leurs sorties à zéro. Cela force le réseau à apprendre des représentations redondantes des données, ce qui empêche une dépendance trop forte vis-à-vis de nœuds spécifiques et encourage

l'apprentissage de caractéristiques plus robustes et généralisables. Pendant les tests ou l'inférence, le dropout est désactivé et le réseau complet est utilisé.

5. **Normalisation en lots (Batch normalization)** : La normalisation en lots est une technique qui normalise les valeurs de sortie des couches intermédiaires d'un réseau de neurones, afin de stabiliser et accélérer l'apprentissage. Elle permet de résoudre les problèmes de saturation de la fonction d'activation et de prévenir les problèmes liés aux écarts de distribution des données d'entrée, assurant ainsi un bon fonctionnement des réseaux de neurones profonds.

6. **Deep CNNs avec skip connections** : Les "skip connections" (connexions sautées) sont des connexions spéciales utilisées dans les réseaux de neurones convolutifs profonds pour relier les couches du réseau de manière alternative lors de la rétropropagation. Elles établissent des liens entre des couches antérieures et des couches ultérieures du réseau, ce qui permet de préserver la propagation d'informations importantes à travers les réseaux profonds. Ces connexions permettent de créer des modèles efficaces avec une faible complexité en termes d'espace et de temps.

   Les architectures de réseau les plus connues utilisent toutes des skip connections pour capturer des caractéristiques à différentes échelles et niveaux d'abstraction, améliorant ainsi leur capacité à apprendre et à représenter les données de manière plus précise. Quelques exemples d'architectures de réseau utilisant des skip connections sont mentionnés.

4.9.1.2 Diverses architectures des CNN

1. **ResNets50**

   Le modèle ResNet50, issu de l'article "**Deep Residual Learning for Image Recognition**" [33] de l'équipe de recherche de **Microsoft**, propose une approche novatrice et élégante. Il ajoute des **skip de connexions** à un réseau de neurones convolutifs profond classique, permettant de contourner plusieurs couches convolutives à la fois.

   Les skip de connexions créent des blocs résiduels, où la sortie des couches convolutives . Le modèle ResNet50 est composé de 50 couches de blocs similaires avec des skip de connexions. Ces connexions permettent de réduire la complexité computationnelle tout en fournissant des combinaisons riches de caractéristiques.

Le modèle ResNet50 comprend une couche convolutive suivie d'une couche de normalisation par lot (batch normalization). Il comporte également deux couches de pooling et un total de 16 modules résiduels. Les modules résiduels sont alternés entre ceux ayant 4 couches convolutives et ceux ayant 3 couches convolutives, chaque couche étant suivie d'une normalisation par lot.

Le modèle a démontré son excellence en atteignant une faible erreur de classification sur le Data Set ImageNet.

2. **U-Net**

   Le U-Net utilise des couches de convolution et des smip de connexions pour créer un modèle capable de capturer à la fois les caractéristiques globales et les détails locaux. Cette architecture est particulièrement adaptée aux tâches de segmentation d'images où une localisation précise des objets est requise.

3. **DenseNet**

   Les connexions denses de DenseNet permettent une communication complète entre toutes les couches du réseau ; utilise une approche unique en connectant chaque couche de convolution à toutes les autres couches de manière séquentielle. Contrairement aux blocs résiduels traditionnels , cela signifie que la sortie de chaque couche est fusionnée avec les feature map de toutes les couches précédentes, créant ainsi des connexions denses. Ces caractéristiques font de DenseNet une architecture efficace et performante dans de nombreuses tâches de vision par ordinateur telles que la reconnaissance d'images et la segmentation.

4. **VGG16**

   VGG16, une version du réseau Very Deep Convolutional Network (VGG), a été développé par des chercheurs de l'Université d'Oxford. Connu pour sa simplicité, VGG16 est considéré comme l'un des meilleurs modèles de la famille VGG.

   L'architecture de VGG16 est à la fois profonde et simple. Elle se compose principalement de couches de convolution et de couches de dropout, qui sont alternées dans le réseau. Une caractéristique distinctive de VGG16 est l'utilisation de multiples filtres de petite taille (3x3) dans chaque couche de convolution. Ces filtres sont ensuite combinés en séquence pour simuler des champs récepteurs plus larges, ce qui permet d'obtenir des représentations plus riches et complexes des images[7].

Malgré sa simplicité, l'architecture de VGG16 présente certains inconvénients. En raison du grand nombre de filtres utilisés, le modèle nécessite une capacité mémoire plus importante et demande des ressources de calcul élevées. Cela se traduit par des temps d'entraînement et d'inférence plus longs et un coût computationnel plus élevé.

Plus spécifiquement, VGG16 est composé de 16 couches de convolution et 5 couches de pooling. Ce modèle a obtenu une précision d'erreur top-5 de 9,9% sur le jeu de données ImageNet, ce qui en fait un choix solide pour les tâches de classification d'images.

5. **Inception v3**

   L'équipe de recherche de Google, sous la direction de Christian Szegedy, s'est principalement focalisée sur la réduction de la charge de calcul des CNN tout en maintenant des performances élevées. Ils ont introduit un module novateur appelé "module Inception", qui se compose principalement de quatre chemins parallèles utilisant des filtres de convolution de tailles 1x1, 3x3 et 5x5. Au fil des années, l'équipe de recherche a proposé plusieurs modèles de plus en plus complexes. L'un de ces modèles, l'Inception v3 , a été introduit à peu près en même temps que ResNet. Ce réseau a été conçu en utilisant des principes de conception novateurs, tels que l'utilisation de convolutions 3x3 plutôt que des convolutions 5x5 ou 7x7 dans les modules Inception. De plus, l'élargissement de la largeur à chaque couche permet d'augmenter la combinaison des caractéristiques pour la couche suivante. L'objectif était de créer un réseau équilibré en termes de complexité de calcul, en trouvant un juste équilibre entre la profondeur et la largeur du réseau[7].

   Le **Figure 4.7** représente différentes architectures et mises en œuvre de modèles utilisant le skip connection , il y a un aperçu des différentes couches de chaque modèle ; qui a été expliqué en détail ci-dessus, chaque couche étant représentée avec une couleur différente .

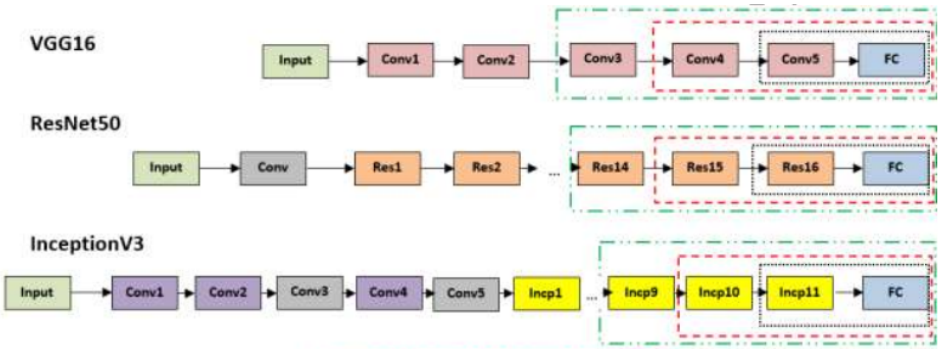


FIGURE 4.7 – Différentes architectures CNN [7]

4.9.2 Construction du modèle

Dans cette partie, Nous avons construit un réseau neuronal profond basé sur `ResNet50v2` avec une couche Dense finale pour classer les images en deux catégories :cancéreuse et non cancéreuse. Le code source disponible sur nos Github : **https://github.com/Nou ...**

Voici une explication plus détaillée des différentes étapes de la construction du nos modèle (ou Algorithme 1) :

1. **Importation de la classe ResNet50V2 de Keras et initialisation du modèle avec les poids pré-entraînés sur ImageNet :** Nous importons la classe ResNet50V2 de la bibliothèque Keras et initialisons le modèle en utilisant les poids pré-entraînés sur ImageNet en spécifiant weights='imagenet'. Cela signifie que nous utilisons des poids qui ont déjà été entraînés sur un vaste ensemble de données d'images appelé ImageNet.
2. **Spécification de la taille d'image d'entrée :**Nous définissons la taille d'image d'entrée à 512x512 pixels avec trois canaux de couleur (RGB). Cela signifie que les images d'entrée doivent avoir ces dimensions pour être compatibles avec le modèle.
3. **Gel de toutes les couches du réseau pré-entraîné :**Nous gelons toutes les couches du réseau pré-entraîné en itérant sur chaque couche à l'aide d'une boucle for et en fixant layer.trainable = False pour chaque couche. Cela permet de conserver les poids pré-entraînés sans les modifier.
4. **Création d'un nouveau modèle Sequential()** : Nous créons un nouveau modèle vide en utilisant la classe Sequential() de Keras. Ce modèle sera utilisé pour ajouter

les nouvelles couches personnalisées au-dessus du réseau ResNet50V2 pré-entraîné.

5. **Ajout du réseau ResNet50V2 pré-entraîné au modèle :** La première couche que nous ajoutons au modèle est le réseau ResNet50V2 pré-entraîné qui a été gelé lors des phases précédentes. Cela nous permet d'utiliser ce réseau comme extracteur de caractéristiques pour les images d'entrée.
6. **Ajout d'une couche GlobalAveragePooling2D() :** Nous ajoutons une couche GlobalAveragePooling2D() qui réduit la dimensionnalité des caractéristiques extraites par le base_model tout en conservant leur information globale. Cette couche calcule la moyenne des valeurs de chaque canal de caractéristiques, ce qui donne une représentation unidimensionnelle de l'image.
7. **Ajout des couches dense :** Nous ajoutons deux couches dense personnalisées au modèle. La première couche dense contient 128 neurones avec une activation ReLU. La deuxième couche ajoutée est une couche Dropout(0,20), qui désactive aléatoirement certaines connexions entre les neurones avec une probabilité de 0,20. Cela aide à éviter le surapprentissage en réduisant la corrélation entre les neurones et en améliorant la généralisation du modèle.
8. **Ajout de la dernière couche de sortie :** La dernière couche ajoutée au modèle est une couche dense avec une seule sortie. Étant donné qu'il s'agit d'un problème de classification binaire (maligne ou bénigne), la fonction d'activation utilisée est "sigmoid". Cela permet de produire une sortie entre 0 et 1, représentant la probabilité d'appartenir à la classe positive.

Algorithm 1: Réseau neuronal profond basé sur ResNet50v2

```
base_model = ResNet50V2(weights='imagenet', input_shape=(512, 512, 3),
    include_top=False)

for layer in base_model.layers:
    layer.trainable = False
model = Sequential()
model.add(base_model)
model.add(GlobalAveragePooling2D())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer="adam", loss='binary_crossentropy', metrics=["
    accuracy"])
```

4.10 Mise en œuvre de l'analyse :

La mise en œuvre de l'analyse implique plusieurs étapes, notamment :

4.10.1 Outils d'implémentation

L'implémentation d'applications basées sur le traitement des données d'images peut être réalisée à l'aide de divers outils et frameworks. Voici quelques-uns des outils couramment utilisés dans ce domaine :

4.10.1.1 Colab

Google Colaboratory est une plateforme de notebooks Jupyter en ligne qui permet aux utilisateurs d'exécuter du code Python dans leur navigateur sans installation requise. Il offre des ressources matérielles gratuites telles que :

— Un espace de stockage temporaire de taille $\geq$70 gigaoctets.

— Différents processeurs de type **CPU,GPU,TPU**.

— Une RAM de taille $\geq$12 gigaoctets.

Ainsi que des bibliothèques standard et couramment utilisées pour l'apprentissage machine comme TensorFlow ou Keras.

4.10.1.2 Kaggle

Kaggle est une plateforme en ligne populaire pour la science des données et les compétitions de machine learning. Les compétitions de Kaggle consistent généralement à résoudre des problèmes de science des données en utilisant un ensemble de données fourni. Les participants soumettent ensuite leurs solutions sous la forme de modèles de machine learning.

4.10.1.3 Python

Python est un langage de programmation interprété, orienté objet et puissant qui permet de développer une grande variété d'applications. Il est facile à apprendre grâce à sa syntaxe simple et claire, ainsi qu'à son large éventail de bibliothèques qui offrent des fonctionnalités supplémentaires pour le développement.
Python peut être utilisé pour diverses applications telles que l'analyse de données, IA,etc. Il a été adopté par plusieurs grandes entreprises technologiques telles que Google ou Facebook, ce qui renforce davantage sa place dans le monde du développement logiciel.

4.10.1.4 Différentes bibliothèques utilisées

voici une description des bibliothèques utilisées dans nos code :

1. **TensorFlow**

   Cette bibliothèque est l'une des plus importantes pour le deep learning en général et pour TensorFlow en particulier. Elle fournit divers outils tels que les couches de réseau neuronal, les fonctions d'activation et les optimiseurs nécessaires à la création et à l'entraînement de modèles.

2. **Keras**

   Keras est une bibliothèque open source qui fournit une interface conviviale pour la création et l'entraînement de modèles de deep learning. Elle s'appuie sur TensorFlow en tant que backend et facilite la construction de réseaux de neurones en fournissant des abstractions simples et intuitives.

3. **NumPy**

NumPy est une bibliothèque fondamentale pour le calcul scientifique en Python. Elle fournit des structures de données efficaces pour la manipulation de tableaux multidimensionnels, ainsi que des fonctions mathématiques pour effectuer des opérations numériques [1].

4. **Pandas**

   Pandas est une bibliothèque populaire pour la manipulation et l'analyse des données en Python. Elle offre des structures de données puissantes, notamment les DataFrames, qui permettent de manipuler et d'analyser facilement des ensembles de données tabulaires [1].

5. **Matplotlib**

   Matplotlib est une bibliothèque de visualisation de données en Python. Elle offre une grande flexibilité pour créer une grande variété de graphiques et de visualisations, tels que des graphiques linéaires, des diagrammes en barres, des histogrammes, des nuages de points, etc. Matplotlib.pyplot est la sous-bibliothèque spécifique de Matplotlib qui est couramment utilisée pour créer des tracés dans le code [1].

6. **Sklearn (scikit-learn)**

   sklearn est une bibliothèque populaire pour l'apprentissage automatique en Python. Elle offre une large gamme d'algorithmes d'ML, de prétraitement des données, d'évaluation des modèles et d'outils pour la sélection de modèles. La sous-bibliothèque sklearn.metrics, mentionnée précédemment, fournit des métriques couramment utilisées pour évaluer la performance des modèles d'apprentissage automatique [1].

7. **Cv2**

   OpenCV est une bibliothèque d'analyse d'image très populaire pour le traitement numérique des images. Elle permet notamment de redimensionner les images ou encore de détecter les contours.

8. **os**

   La bibliothèque OS fournit un moyen simple et portable d'utiliser certaines fonctionnalités du système d'exploitation telles que la création/suppression de fichiers, la modification des variables environnementales, etc.

4.10.2 Entraînement du modèle

Nous avons entraîné le modèle sur leurs DataSets traités avec différentes configurations d'hyperparamètres ,Les poids sont ajustés pendant l'apprentissage afin que le modèle puisse apprendre à catégoriser les différentes classes avec une certaine précision . La Figure 4,9 représente résultat d'entraînement du modéle

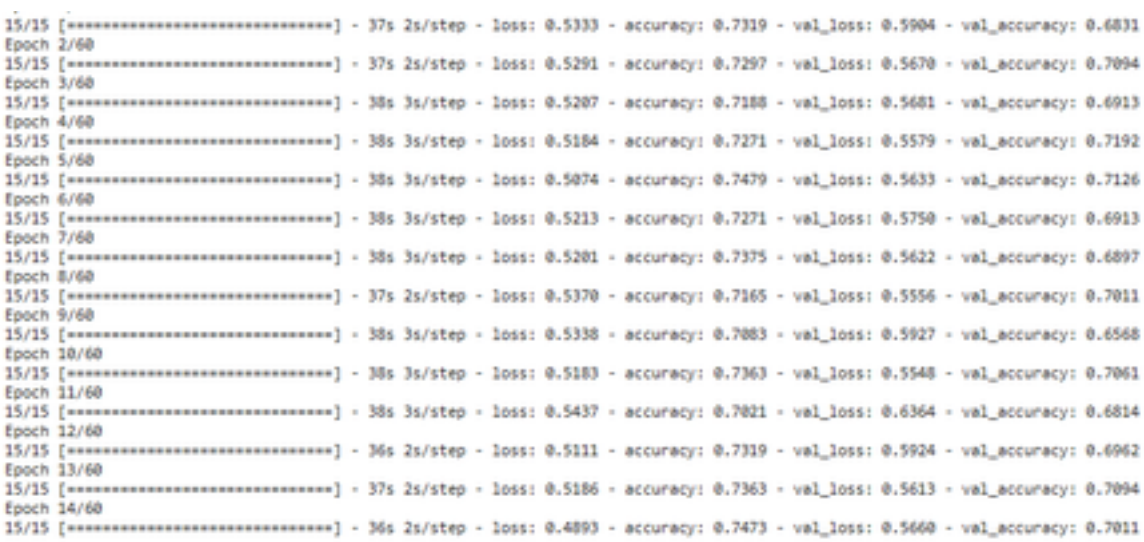


FIGURE 4.8 – Entraînement du modèle.

4.11 Visualisation du Accuracy et Loss

Nous avons utilisées les données d'accuracy et de loss stockées dans l'historique de l'entraînement du modèle pour tracer les courbes d'accuracy et de loss à travers les epochs qu'ils sont présente dans la Figure 4.9 . Cela permet de visualiser comment ces métriques évoluent pendant l'entraînement et la validation du modèle.

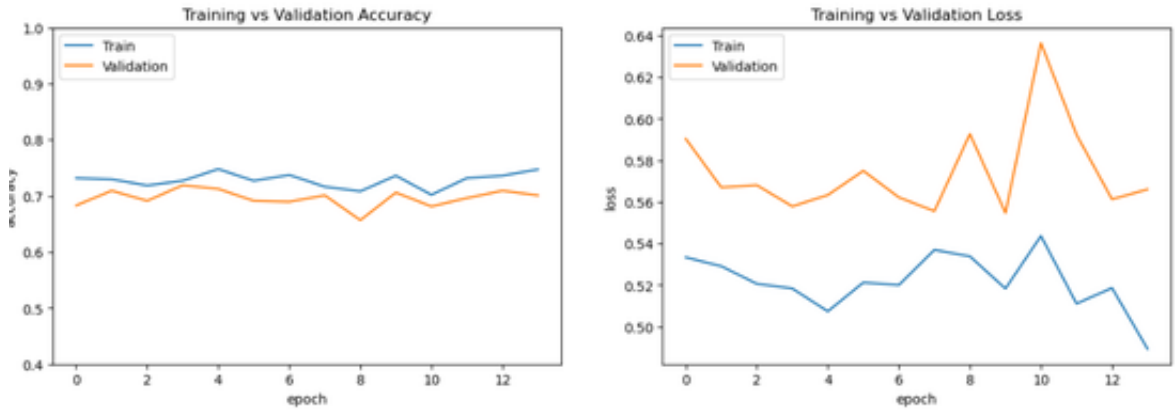


FIGURE 4.9 – Visualisation du Accuracy et Loss

4.12 Interprétation et communication des résultats :

La Interprétation et communication des résultats sont des étapes cruciales dans ML. Nous avons effectué ces tâches de manière approfondie. Voici une explication plus détaillée :

4.12.1 Prédiction

Afin d'évaluer la qualité du modèle entraîné, les résultats ont été prédits à partir des données de test (validation) et comparés aux valeurs observées. Le seuil a été fixé à 0.5 et les valeurs prédites supérieures à 0.5 ont été considérées comme positives (1).

La Figure 4.10 représente La prédiction de nos modèle : Selon les résultats de vos prédictions, vous avez obtenu une liste ypred contenant des prédictions binaires pour chaque image. Une valeur de 1 indique que l'image est prédite comme cancéreuse et une valeur de 0 indique qu'elle est prédite comme normale. La résultat de nos prédiction représente est le suivant :

— Les images prédit cancéreuse est : **3**13

— Les images prédit normale est : **2**96

```
#Prédiction des résultats en utilisant le modèle sur le générateur de validation.
pred = model.predict(validation_generator)
```

```
40/40 [==============================] - 9s 206ms/step
```

```
# prediction by the AI
pred
```

```
array([[0.30161956],
       [0.59507996],
       [0.4412763 ],
       [0.3047006 ],
       [0.6544019 ],
       [0.63904023],
       [0.29479995],
       [0.6174365 ],
       [0.43504983],
       [0.65982604],
       [0.5224252 ],
       [0.67800605],
       [0.7196614 ],
       [0.7073742 ],
       [0.5664427 ],
       [0.6105107 ],
       [0.4963043 ],
       [0.467375671,
```

FIGURE 4.10 – Prédiction

4.13 Évaluation des résultats :

L'évaluation des résultats peut être complexe et dépendante du contexte, mais voici quelques paramètres couramment utilisés pour évaluer :

1. **Matrice de confusion**

   Une matrice de confusion est une méthode courante pour évaluer les performances d'un modèle de classification. Elle permet de visualiser le nombre de prédictions correctes et incorrectes du modèle en comparant les valeurs réelles avec celles prédites. Dans le cas d'une classification binaire (comme nos modèle), cette matrice sera constitué 4 éléments :

   1.1 **True Positive (TP)** : observation positive classée comme positive.

   1.2 **False Negative (FN)** : observation positive classée comme négative.

   1.3 **False Positive (FP)** :observation négative classée comme positive.

   1.4 **True Negative(TN)** :observation négative classée comme negative.

La Figure 4.11 représente la Matrice de confusion après d'exécutions

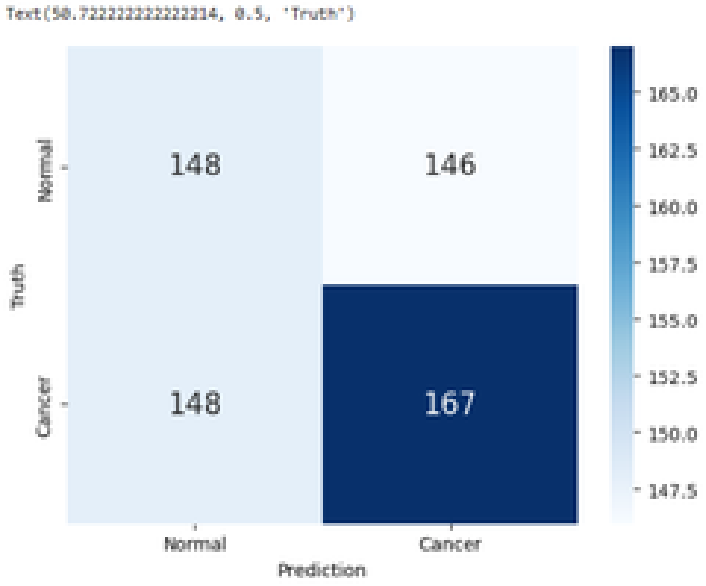


FIGURE 4.11 – Matrice de confusion

2. **Rapport de classification**

   Un rapport de classification est un moyen d'évaluer les performances d'un modèle de classification en fournissant des informations sur la précision, le rappel, le score F1 et le support pour chaque classe.

   Le rapport de classification peut être généré à partir de la matrice de confusion. Il affiche les résultats pour chaque classe ainsi que leur moyenne pondérée ou non-pondérée selon l'option choisie.

   La Figure 4.12 représente le rapport de classification après Nous avons deux classes : "normal" et "cancer". Pour chacune des classes, nous avons calculé la précision ,le rappel et le score F1. La moyenne pondérée des métriques est également fournie sous forme d'une ligne supplémentaire dans ce tableau.Le poids est déterminé par le nombre total d'exemples dans chaque classe . Enfin ,la colonne support indique combien d'échantillons sont présents pour chacune des deux classes.

	precision	recall	f1-score	support
0	0.50	0.50	0.50	294
1	0.53	0.53	0.53	315
accuracy			0.52	609
macro avg	0.52	0.52	0.52	609
weighted avg	0.52	0.52	0.52	609

FIGURE 4.12 – Rapport de classification

3. **Temps de traitement :** Le temps d'entraînement du modèle a été mesuré à environ 1 heure sur un GPU NVIDIA Tesla P100. Cependant, le temps de traitement peut varier en fonction des ressources matérielles disponibles et de la taille des données.
4. **Efficacité en mémoire :** la taille des images utilisées (512 x 512) peut nécessiter une grande quantité de mémoire.
5. **Scalabilité :** Il n'y avait pas suffisamment d'informations disponibles pour évaluer la scalabilité du modèle. parceque , il faut achetée plus des ressources et cherche plus des ensembles des données , cette action est trées coûteuses et rares à obtenir.
6. **Extensibilité :** Le code fourni est potentiellement extensible pour inclure plus de classes ou effectuer une classification multiclasse en modifiant quelques paramètres.
7. **Facilité d'utilisation :** Nos code fournit des explications détaillées sur les étapes à suivre pour entraîner et évaluer le modèle, mais une connaissance préalable de Python et TensorFlow est requise.
8. **Coût :** Il n'y a pas ou peu de coûts directs associés à ce projet car il utilise un environnement gratuit en ligne offert par Kaggle ou Colab avec accès à un GPU pour l'enrainement du modele.
9. **Complexité algorithmique :**
   - 9.1 **Complexité temporelle** : L'utilisation de ResNet50V2 implique que le temps d'exécution du modèle peut être relativement long en fonction de la taille des données d'entrée.
   - 9.2 **Complexité spatiale** : La complexité spatiale est également importante car le traitement d'un grand nombre d'images simultanément nécessite une grande quantité de mémoire supplémentaire.

4.14 Discutions et Résultat

Nous avons réussi à construire un classificateur performant qui peut discriminer avec précision les images mammographiques malignes/bénignes en utilisant des techniques de Deep Learning, notamment le modèle ResNet50V2, combiné à des prétraitements appropriés tels que la normalisation des données.
Le modèle a été entraîné sur un ensemble de données d'images mammographiques prétraitées, atteignant une précision moyenne de 70%.
Les résultats obtenus démontrent que l'utilisation de ces techniques a encore amélioré les performances du modèle. En conclusion, ce travail met en évidence comment les architectures de Deep Learning telles que ResNet peuvent être efficacement utilisées pour résoudre des problèmes complexes tels que la classification d'images.
Cela a été rendu possible grâce à l'utilisation judicieuse d'outils tels que TensorFlow et Keras, qui non seulement accélèrent notre travail, mais aussi facilitent grandement la création, la formation et l'évaluation de modèles performants pour les analystes.

4.15 Conclusion

En résumé, ce travail constitue une contribution significative dans le domaine de la détection du cancer du sein en utilisant des techniques de Deep Learning et l'analyse de Big Data. Nous avons développé un modèle basé sur ResNet50V2 qui a obtenu de bons résultats en termes de précision et de rappel. L'ensemble du processus, depuis la collecte des données jusqu'à l'évaluation du modèle, nous a permis de traiter efficacement les images médicales et les données cliniques, améliorant ainsi la détection du cancer du sein.
Ces résultats encourageants ouvrent de nouvelles perspectives pour des améliorations futures et l'application de cette approche dans un contexte clinique réel, afin d'aider les médecins dans leur diagnostic précoce et leur prise de décision.

Chapitre 5

Conclusion générale

5.1 Contributions

Ce projet de recherche avait pour objectif le développement d'une méthode de détection du cancer du sein en utilisant une analyse approfondie des Big Data. Notre contribution principale réside dans l'exploration complète des différentes étapes du traitement des Big Data, en mettant l'accent sur l'utilisation d'algorithmes de Deep Learning basés sur les réseaux de neurones à convolution (CNN).

En se basant sur les recherches antérieures et les principes des CNN, nous avons développé une méthode spécifique pour la détection du cancer du sein. Notre approche s'est appuyée sur une analyse approfondie des données volumineuses disponibles, principalement des images médicales, et l'utilisation des modèles ResNet50V2 et VGG16 pour la classification des images médicales liées au cancer du sein.

5.2 Limites et critiques

Malgré nos contributions, certaines limites et critiques doivent être soulignées. Nous avons principalement exploité des images médicales comme source de données, mais il est possible que d'autres sources pertinentes, telles que les données génomiques, n'aient pas été prises en compte. L'inclusion de ces données aurait pu fournir des informations précieuses sur les variations génétiques associées au cancer du sein.

Bien que nos modèles aient montré des résultats prometteurs, des défis persistent. Des recherches supplémentaires sur les techniques d'optimisation des algorithmes de deep

learning sont nécessaires pour réduire les faux positifs et les faux négatifs.
La gestion et l'analyse des données ont été confrontées à des limitations significatives en termes de stockage et de capacité de traitement. Malgré l'utilisation d'outils avancés tels que Hadoop et Spark, nos machines n'avaient pas les ressources nécessaires pour effectuer des analyses de données efficaces.
De plus, la quantité de données utilisée dans notre étude peut également avoir un impact sur la scalabilité de notre méthode. L'augmentation du volume de données peut entraîner des temps de traitement plus longs et une demande accrue en ressources. Ces aspects doivent être pris en compte lors de l'application de notre méthode à des ensembles de données plus vastes.

5.3 Perspectives futures

Pour les travaux futurs, il est essentiel d'explorer l'intégration d'autres sources de données, d'améliorer les performances des modèles existants et d'étendre la méthode de détection à d'autres types de cancer ou à d'autres domaines. L'utilisation de l'intelligence artificielle et de l'analyse des Big Data présente un énorme potentiel pour améliorer le diagnostic et le traitement du cancer. Des recherches supplémentaires dans ce domaine sont nécessaires pour faire progresser la lutte contre cette maladie.

Références

[1] Guillaume Saint-Cirgue. *Apprendre Machine Learning dans une semaine*. 2019.

[2] L. R. Nair, S. D. Shetty, and S. D. Shetty. Applying spark based machine learning model on streaming big data for health status prediction. *Computers & Electrical Engineering*, 2018.

[3] G. Asencio–Cortés, A. Morales–Esteban, X. Shang, and F. Martínez–Álvarez. Earthquake prediction in california using regression algorithms and cloud-based big data infrastructure. *Computers Geosciences*, 115 :198–210, 2018.

[4] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh.

[5] Makrufa Hajirahimova and Aybeniz Aliyeva. A survey on deep learning in big data analytics. *Institute of Information Technology - Azerbaijan National Academy of Sciences*, 2020.

[6] Amir Masoud Rahmani and et al. Artificial intelligence approaches and mechanisms for big data analytics : a systematic study. *PeerJ Computer Science*, 7 :e488, 2021.

[7] Deep convolutional neural networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, 157 :19–30, 2018.

[8] Mozamel M. Saeed, Zaher Al Aghbari, and Mohammed Alsharidah. Big data clustering techniques based on spark : a literature review. 6 :e321.

[9] Richa Gupta. Journey from data mining to web mining to big data. *arXiv preprint arXiv :1404.4140*, 2014.

[10] Pirmin Lemberger. *Big Data et Machine Learning - Manuel du data scientist*. 2016.

[11] Sanjay Ghemawat Jeffrey Dean. Mapreduce : Simplified data processing on large clusters. *Google, Inc.*, 2004.

[12] Yuri Demchenko, Cees De Laat, and Peter Membrey. Defining architecture components of the big data ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2014.

[13] Judith S. Hurwitz, Alan Nugent, Fern Halper, and Marcia Kaufman. *Big data for dummies*. John Wiley & Sons, 2013.

[14] Adanma Cecilia Eberendu. Unstructured data : an overview of the data of big data. *International Journal of Computer Trends and Technology*.

[15] Ronen Feldman and James Sanger. *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[16] Khelifi Hakima. *cour chapitre 2 Big Data*. 2022.

[17] Stuart Russell and Peter Norvig. *Intelligence artificielle*. 3e edition, 2010.

[18] Stuart Russell and Peter Norvig. *Artificial Intelligence : A Modern Approach*. Prentice Hall, Hoboken, 4th edition, 2020.

[19] Amir Masoud Rahmani, Ameen Alsalem, Mamoun Alazab, Muhammad Asim Noor, Naif Radi Aljohani, and BB Gupta. Artificial intelligence approaches and mechanisms for big data analytics : a systematic study. *PeerJ Computer Science*, 7, 2021.

[20] A. A. AlZubi. Big data analytic diabetics using map reduce and classification techniques. *The Journal of Supercomputing*, 2020.

[21] Shweta Mittal and Om Prakash Sangwan. Big data analytics using machine learning techniques. In *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 203–207, 2019.

[22] M. Ianni, E. Masciari, G. M. Mazzeo, M. Mezzanzanica, and C. Zaniolo. Fast and

effective big data exploration by clustering. *Future Generation Computer Systems*, 102 :84–94, 2020.

[23] F. Pulgar-Rubio, A.J. Rivera-Rivas, M.D. Pérez-Godoy, P. González, C.J. Carmona, and M.J. del Jesus. Mefasd-bd : multi-objective evolutionary fuzzy algorithm for subgroup discovery in big data environments-a mapreduce solution. *Knowledge-Based Systems*, 117 :70–78, 2017.

[24] Richard H Ip, Li-Minn Ang, Kelvin P Seng, John C Broster, and James E Pratley. Big data and machine learning for crop protection. *Computers and Electronics in Agriculture*, 151 :376–383, 2018.

[25] Xiaojin Zhu. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, 2005.

[26] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, pages 281–297, Berkeley, CA, 1967. University of California Press.

[27] Simon Haykin. *Neural Networks—A Comprehensive Foundation*. Pearson Education, Singapore, 2003.

[28] Bilal Jan, Haleem Farman, Murad Khan, Muhammad Imran, Ihtesham Ul Islam, Awais Ahmad, Shaukat Ali, and Gwanggil Jeon. Deep learning in big data analytics : A comparative study. *Computers Electrical Engineering*, 2019.

[29] Badr-Addeen Hammou, Ahmed Al Lahcen, and Salah Mouline. Towards a real-time processing framework based on improved distributed recurrent neural network variants with fasttext for social big data analytics. *Information Processing Management*, 2020.

[30] Nadia Frikha and Mehdi Chlif. Un aperçu des facteurs de risque du cancer du sein. *Bulletin de l'Académie Nationale de Médecine*, 205(5) :519–527, 2021.

[31] Mouhoub Ahlem. cour6 cancer de sein, 2022.

[32] Rohit Thanki Khalid Shaikh, Sabitha Krishnan. Artificial intelligence in breast cancer early detection and diagnosis. 2021.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.