République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université 8 Mai 1945 - Guelma - Faculté des Mathématiques,
d'Informatique et des Sciences de la matière

Département d'Informatique

Mémoire de Fin d'études Master

Specialté : Informatique

Option : Systèmes Informatiques

Thème

Traitement prédictif des données manquantes médicales par méthode d'apprentissage

---

Encadré par :

Dr. BENHAMZA Karima

Présenté par :

Melle NAIDJA Hanane

Membres du Jury :

Dr. ABDELMOUMENE Hiba

Dr. BOUGHAREB Djalila

Juin 2023

Dédicace

À mes chers parents et mes petites sœurs adorées, Il est difficile de mettre en mots toute la gratitude et l'amour que je ressens envers vous. Vous avez été mes piliers, ma source d'inspiration et ma force tout au long de ma vie. Cette dédicace est un témoignage de l'immense reconnaissance que j'ai envers vous.

Chers parents, vous êtes les architectes de ma vie, ceux qui m'ont inculqué des valeurs solides et m'ont encouragé à poursuivre mes rêves. Votre amour inconditionnel, votre soutien indéfectible et vos sacrifices désintéressés ont fait de moi la personne que je suis aujourd'hui. Vous avez toujours cru en moi, même lorsque je doutais de moi-même, et vous m'avez montré que rien n'est impossible avec la détermination et le travail acharné.

Mes chères petites sœurs, vous êtes mes rayons de soleil, ma joie de vivre et ma source de bonheur. Vous avez illuminé ma vie de rires, de complicité et de moments précieux. Votre innocence, votre curiosité et votre amour débordant ont égayé chaque jour et ont donné un sens plus profond à mes réalisations. Je suis fière d'être votre grande sœur, et je serai toujours là pour vous, prête à vous soutenir dans vos propres chemins.

À ma chère amie Nassima, Il n'y a pas de mots assez forts pour exprimer à quel point ta présence dans ma vie est précieuse. Cette dédicace est un humble témoignage de mon amour et de ma reconnaissance envers toi, ma meilleure amie. Depuis le premier jour où nos chemins se sont croisés, tu as illuminé ma vie de ta gentillesse, de ta loyauté et de ton soutien inconditionnel. Tu es celle sur qui je peux toujours compter, celle avec qui je partage mes joies et mes peines, mes rêves et mes craintes. Tu as été mon roc dans les moments difficiles, et ma complice dans les instants de bonheur.

Hanane

Remerciement

Alhamdoulilah , Je remercie Dieu le tout puissant de m'avoir donné la santé et la volonté d'entamer ce mémoire et de le terminer.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurais pas vu le jour sans l'aide et l'encadrement du Dr Benhamza Karima, Je tiens à exprimer ma profonde gratitude et ma sincère reconnaissance pour votre soutien inestimable tout au long de mon mémoire de fin d'études. Votre expertise, votre patience et votre dévouement et surtout votre gentillesse ont été des éléments clés dans la réussite de ce projet. Je souhaite également souligner l'impact positif que vous avez eu sur ma confiance en moi et ma capacité à relever des défis. Votre encouragement constant et votre soutien inconditionnel m'ont permis de croire en mes capacités et de me dépasser dans ma future carrière, je vous serai éternellement reconnaissante très chère Mme Benhamza.

Mes remerciements s'adressent également aux membres du jury, je tiens à vous remercier pour votre contribution à mon parcours académique. Votre présence en tant que membre du jury a été un honneur pour moi, et j'apprécie profondément votre participation et vos conseils.

J'adressent aussi mes remerciements et à tous mes professeurs pour la qualité de l'enseignement qui m'ont prodigué au cours de mes années passées à l'université. Je remercie toute personne qui a contribué d'une manière ou d'une autre à la réalisation de ce mémoire.

Résumé

Après l'explosion des données dans le monde ces dernières années, tous les domaines ont été envahi par la technologie « Big Data » et se sont retrouver face à ses défis. Le domaine médical n'a pas eu d'exception et s'est retrouver aussi face avec un plus grand défi : le problème des données manquantes ou « Missing Data ».

Dans ce travail, on s'intéresse à l'analyse du Big Data Medical pour prévoir les tendances et les comportements futures des données avec une grande fiabilité dans le cadre de traitement de données manquantes. Les données manquantes sont très fréquentes dans le domaine médical et engendre malheureusement des difficultés immenses de diagnostic. Leur traitement aussi est très sensible vu que la vie des gens en dépend.

Le but de ce travail est de montrer l'importance des méthodes d'apprentissage dans le traitement des données manquantes dans le domaine médical et de proposer un système intelligent d'imputation de données basé sur les méthodes d'apprentissage profond. Finalement, les résultats, très satisfaisants obtenus de la combinaison de différentes méthodes appliquées sur deux Datasets Médicaux, nous a permet de souligner l'intérêt du modèle proposé.

Mots clés : Big Data Médical, Apprentissage machine, Fuzzy K-Means, Apprentissage profond, Réseau Antagoniste Génératif.

Abstract

After the explosion of data worldwide in recent years, all fields have been invaded by the "Big Data" technology and have faced its challenges. The medical field has been no exception and has faced an even greater challenge : the problem of missing data.

In this work, we focus on the analysis of Medical Big Data to predict future trends and behaviors of data with high reliability in the context of missing data treatment. Missing data is very common in the medical field and unfortunately leads to immense diagnostic difficulties. Their treatment is also very sensitive since people's lives depend on it.

The goal of this work is to demonstrate the importance of learning methods in the treatment of missing data in the medical field and to propose an intelligent data imputation system based on deep learning methods. Finally, the highly satisfactory results obtained from the combination of different methods applied to two Medical Datasets have allowed us to highlight the significance of the proposed model.

keywords : Big Data, Medical Dataset, Analytical Methods, Machine Learning, Deep Learning, Fuzzy K-means, Generative Antagonist Network.

Table des matières

Liste des tableaux

Table des figures

Introduction

On parle depuis quelques années du phénomène "Big Data", qui signifie littéralement "Mégadonnées", "Grosses données" ou encore "Données massives". Ce concept fait référence à un très vaste ensemble de données qu'aucun outil classique de gestion des bases de données ou de gestion de l'information ne peut réellement être appliqué. Cette explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles technologies pour l'analyser. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, le traitement et la présentation des données.

Ainsi de nouveaux défis ont apparu en raison de cette masse d'information générée par l'écosystème informationnel impactant la performance analytique. En effet, les limites se localisent au niveau de l'extraction des sources, de la transformation des données, de temps réponse de la couche analytique, les limitations physiques du réseau et bien d'autres contraintes.

En santé, comme dans plusieurs d'autres domaines, les progrès technologiques ont fait exploser la quantité d'informations recueillies à chaque instant. La récupération des données massives est relativement facile, contrairement à son traitement de façon exacte et efficace en raison de son incomplétude. Ce problème a obligé les chercheurs à relever le défi dans cet axe de recherche difficile qui est le traitement des données manquantes. Elles engendrent malheureusement des difficultés immenses de diagnostic et de prise de décision , leurs traitements dans le domaine médicale est sensible vu que la vie des gens en dépend.

Notre projet de fin d'étude a pour but d'étudier les technologies et les méthodes du « Big Data ». Nous nous intéresserons particulièrement aux méthodes de traitement de données manquantes dans le domaine médical. L'intérêt de cette recherche est d'analyser et de

comparer les méthodes utilisés afin de proposer un nouveau modèle.

Ce mémoire est structuré comme suit :

- Dans le chapitre 1, on présentera la technologie « Big Data » avec ses concepts et ses caractéristiques ainsi que son impact dans le domaine médical .
- Dans le chapitre 2, les méthodes utilisées pour le traitement des données manquantes dans les « Big Data » médical seront détaillées.
- Le chapitre 3 traitera la conception et l'implémentation du modèle proposé pour imputer les données manquantes dans les « Big Data » médicales. La plateforme Spark est utilisée pour le traitement des données et la présentation des résultats obtenus.

  Finalement, ce mémoire est clôturé par une conclusion générale, des perspectives de ce travail et des références bibliographiques utilisée.

Chapitre 1

BIG DATA

1.1 Introduction

L'utilisation de l'Intelligence Artificielle(IA) est grandissante dans tous les domaines existants et le domaine de la santé ne fait pas exception. Les outils de l'IA, telles les techniques d'apprentissage automatique, épaulent les professionnels dans leur démarche pour assurer des services rapides et de qualité. En effet, ils révolutionnent la pratique médicale par l'allocation des ressources au diagnostic des maladies complexes, en mobilisant une immense quantité d'informations communément appelé "BIG DATA". Cependant, ce dernier est confronté à de nombreux défis.

Dans ce chapitre, nous aborderons les bases du « Big Data », son rôle et ses caractéristiques dans le domaine médical.

1.2 Définitions

Le Big Data est un flux massif de données dans un format structuré, non structuré ou hétérogène qui a été accumulé en raison de l'augmentation exponentielle et continue du volume des données saisies par les organisations, telles que les médias sociaux, le gouvernement, l'industrie et la science [2] [6]. Littéralement, ce sont des grosses données ou volume massif de données,on parle aussi de "Data masse" par similitude avec la biomasse,et conceptuellement ce terme vulgarise à la fois la représentation du volume des données mais aussi les infrastructures liées au traitement de ces données [7]. Le Big Data est aussi défini comme étant un ensemble de données numériques produites par l'utilisation de nouvelles technologies, qui est difficile à traiter et à analyser en utilisant les outils classiques de gestion de base de données.

1.3 Caractéristiques du Big Data

Les caractéristiques fondamentales du Big Data ont été définies à l'origine au début des années 2000 par le modèle 3V, qui comprend le Volume, la Variété et la Vélocité. Ce modèle a depuis été étendu à 7V comme présenter dans la figure 1.2, ajoutant la variabilité, la véracité, la valeur et enfin la visualisation [1] [8].

FIGURE 1.1 – Modèle 7V du Big Data [1]

1.3.1 Volume

Le volume est une composante primordiale du Big Data et représente principalement la relation entre la taille et la capacité de traitement. Cette variable évolue rapidement à mesure de l'accumulation des données.
Dans les systèmes d'information mis en place dans les entreprises, les volumes de données traitées se mesurent en Téraoctets. Le challenge immédiat est d'être en capacité de traiter des Pétaoctets et bientôt des Exaoctets puis des Zettaoctets. Ce qui nécessite de développer un plan pour gérer la quantité immense de données qui sera mis en jeu et prévoir son hébergement. Comme le révèlent les prévisions, le volume de données générées dans le monde devrait dépasser 180 zettaoctets à l'horizon 2025 [9].

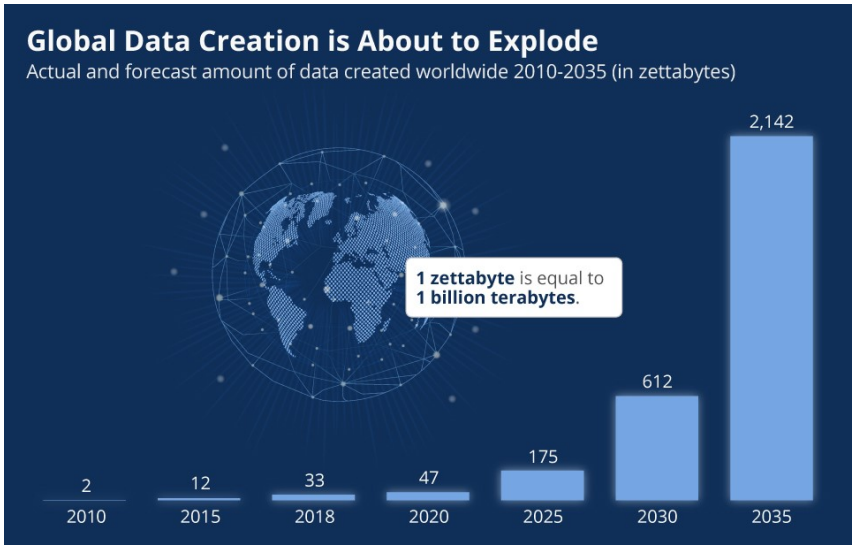

FIGURE 1.2 – Volume mondial annuel de données [2]

1.3.2 Variété

La variété décrit la grande diversité d'information contenue qui doit faire l'objet d'une analyse collective [6]. De nouveaux types de données issues des réseaux sociaux et d'appareils connectés, entre autres, complètent les types d'informations structurées existantes. Par exemple : fichiers médicaux, documents graphiques, documents Web, cartes bonus et comportement de recherche sur Internet. Les données non structurées telles que la voix et les médias sociaux rendent difficiles leur traitement et leur catégorisation de ces derniers.

1.3.3 Vélocité

La vélocité fait référence à la grande vitesse d'accumulation des données. Dans le Big Data, la vélocité détermine le potentiel des données i.e. A quelle vitesse les données sont générées et traitées pour répondre aux demandes [10]. Le flux de données est souvent vaste et continu, nécessitant des plates-formes et des capacités capables non seulement de gérer ces volumes importants, mais aussi de traiter ce flux en temps réel [10].

1.3.4 Variabilité

La variabilité définie à quelle vitesse et fréquence la structure des données change. L'important est d'établir si la structure contextuelle du flux de données est régulière et fiable même dans des conditions d'imprévisibilité extrêmes. La définition de la variabilité des données permet de gérer les données de manière à les structurer,même dans des environnements imprévisibles et variables en tenant compte de toutes les circonstances possibles[11].

1.3.5 Véracité

La véracité permet de vérifier la qualité et l'origine de l'information. Cela fait référence aux incohérences et à l'incertitude des données. Cependant, établir la confiance en l'information représente un challenge de taille des Big Data qui se traduit par l'importance de traiter et de gérer l'incertitude inhérente liée à certains types de données [12].

1.3.6 Valeur

Ce concept décrit la valeur à obtenir à partir des données et comment les mégas données obtiennent de meilleurs résultats à partir de données stockées. Être capable de tirer de la valeur des Big Data est une condition préalable, car la valeur des Big Data augmente considérablement en fonction des informations qui peuvent en être tirées [12].
Les données sont variables en qualité et en volume. Leur transformation et analyse sont avant tout dans le but de créer de la valeur. Le degré de cette valeur dépendra de la pertinence (véracité) et du service que l'entreprise mettra en œuvre pour analyser et adresser le résultat obtenu.

1.3.7 Visualisation

La visualisation est le dernier point des projets en mégadonnées. Il s'agit, littéralement, de visualiser l'information afin de tirer profit de l'œil humain dans l'analyse des données. Cette visualisation peut se faire de plusieurs manières : des graphes 2D ou 3D, des réseaux, des cartes, etc. L'objectif est de rendre l'information visuellement exploitable et lisible pour permettre la prise de décision [13].

1.4 Big Data médical

Le Big Data s'applique à beaucoup de domaines, mais l'un des plus touchés par cette technologie révolutionnaire est le domaine médical. Dans ce domaine, les Big Data (ou données massives) correspondent à l'ensemble des données disponibles sur la santé recueillies par différentes sources [14].
Les techniques médicaux et les activités cliniques sont extrêmement complexes et diversifiés. La collecte de données dans les Big Data et les modèles appliqués deviennent plus complexes ou de nouveaux appareils portables, sources de données et applications mobiles utilisés par les patients augmentent énormément. Cette croissance exponentielle des données va permettre aux médecins d'améliorer leurs diagnostics par une meilleure compréhension de l'état des patients. Les signes impressionnants de progrès dans le Big Data médical vont conduire à une réévaluation de la vision mondiale des systèmes et des organisations de santé [15].

1.5 Caractéristiques du Big Data médical

Le « Big Data » dans le domaine de la santé a ses propres caractéristiques en plus des caractéristiques '7V' du Big Data précédemment citées. Ces dernières sont l'hétérogénéité, l'incomplétude, l'actualité et l'éthique médicale.

1.5.1 Hétérogénéité

Le « Big Data » médical contient des données de nature différente avec souvent des formats incompatibles. Il est difficile de regrouper des sources aussi variées, mais critiques d'informations dans un format de données intuitif ou unifié. Ces données peuvent être classées en données structurées, non structurées et semi-structurées.
Cependant, la majorité des données médicales sont non structurées, les options de saisie structurées peuvent ne pas suffire à capturer des données de nature complexe (données non standards, source ou provenances différentes telles : la tomodensitométrie, de l'IRM, Rayons X, la surveillance Holter, l'angiographie et les laboratoirs [16].

1.5.2 Éthique médicale

L'émergence des Big Data et leurs perspectives prometteuses posent des problèmes nouveaux aux professionnels de santé d'ordre judiciaire, médical et de réparation relatifs à des nouvelles exigences de légitimité du droit à l'information provoquant une certaine désorganisation et bouleversement dans la relation médecin-patient. À cela s'ajoute une véritable prise de conscience et remise en cause éthique sur la confidentialité, le droit et liberté d'accès, la commercialisation, la sécurité, la responsabilité et le secret médical qui entourent l'usage de ces données médicales personnelles via ces Big Data [17].

1.5.3 Actualité des données médicales

Les données de santé ne sont pas statiques et la plupart des éléments nécessitent des mises à jour fréquentes afin de garder leurs pertinences. Pour certains ensembles de données, comme les signes vitaux des patients, ces mises à jour peuvent se produisent toutes les secondes [18].

1.5.4 Incomplétude

Les suivis médicaux des patients de l'hôpital exigent d'enregistrer leurs profils médicaux complets tels les traitements, les cas cliniques, etc. Cette masse d'informations précieuses nécessite des ressources de stockage très coûteuses ce qui conduit éventuellement au problème délicat de l'incomplétude des données dans les datasets [19].

1.6 Utilisation du Big Data dans le domaine médical

L'utilité des Big Data dans le domaine de la santé est inconstatable. Leurs apports fructueux sont décrits ci-dessous : [20]

- **La prévention et la prise en charge des maladies** Les données de santé collectées permettront d'identifier des facteurs de risque pour certaines maladies. Ces facteurs serviront ensuite pour construire des messages de prévention, et mettre en place des programmes à destination des populations à risque.
- **Prédiction des épidémies** Disposer de nombreuses informations sur l'état de santé des individus dans une région donnée permettent de repérer l'augmentation de l'incidence des maladies ou des comportements à risque, et d'alerter les autorités sanitaire.
- **Amélioration de la pharmacovigilance** optimisation de la surveillance et de la prévention des risques d'effets indésirables, les sociétés pharmaceutiques utilisent les informations des Big Data pour créer des modèles et des simulations réalistes afin de tester leurs produits.
- **Gestion de flux des patients** L'objectif est de pouvoir déployer le nombre d'employés adéquats lorsqu'il y a des pics de fréquentation et d'assurer une utilisation optimale des ressources disponibles.

1.7 Données manquante dans les Big Data

Les données manquantes (Plus communément appelées "Missing Data") sont généralement attribuées à une erreur humaine lors du traitement des données. Erreur due au dysfonctionnement du matériel, abandon dans les études et fusion de données non liées.

Ce problème de valeurs manquantes est généralement courant dans tous les domaines qui traitent des grands flux de donnéess et provoque des problèmes différents comme la dégradation des performances, les problèmes d'analyse des données et les résultats biaisés(différences entre les valeurs manquantes et complètes). De plus, la gravité du problème des valeurs manquantes dépend en partie de la quantité de données manquantes, du modèle de données manquantes et du mécanisme sous-jacent à l'absence de ces données [4].

1.7.1 Types de données manquantes

Il existe trois types de données manquantes [4]. :

- **Missing Completely At Random (MCAR)** : Une donnée est MCAR. C'est-à-dire manquante de façon complètement aléatoire, c'est le cas lorsque les observations manquantes ne dépendent pas des mesures observées et non observées. Il n'y a pas de mécanisme caché lié a au données manquantes.
- **Missing At Random (MAR)** : Les données (MAR) ne manquent pas de façon complètement aléatoire. La probabilité d'absence n'est liée qu'à une ou plusieurs autres variables observées, c'est-à-dire que les données observées sont statistiquement liées aux variables manquantes.
- **Missing Not At Random (MNAR)** : Cela fait référence aux données manquantes qui ne sont ni de type MCAR ni de type MAR. Pour ce type, la gestion des valeurs manquantes est généralement impossible, car cela dépend des données invisibles.

1.8 Plateformes Big Data

La manifestation des mégas données a suscité des défis cruciaux pour gérer leur traitement et leur analyse. Ce qui a conduit à la conception et à la mise en place de plates-formes adaptées. L'avantage principal de ces plateformes est de réduire la complexité des données par leur analyse. Nous présentons en dessous les plateforme d'analyses du Big Data les plus utilisés :

1.8.1 Apache Hadoop

Hadoop est un framework logiciel Open Source [21] basé sur le langage Java permettant de stocker des données et de lancer des applications sur des grappes de machines standard. Cette solution a offert un espace de stockage massif pour tous les types de données et aussi la possibilité de prendre en charge une quantité de tâches virtuellement illimitée.

Hadoop repose sur un ensemble de machines formant un cluster Hadoop. Chaque machine est appelée nœud. C'est l'addition des capacités de stockage et traitement de ses nœuds qui lui assure un important système de stockage et une grande puissance de calcul. Le système de stockage est appelé HDFS (Hadoop Distributed File System).

Hadoop utilise "MapReduce" pour traiter rapidement les données massives [20]. La fonctionnalité de "MapReduce" est basée sur le concept du parallélisme. Le modèle "MapReduce" effectue le traitement en deux phases distinctes. Chaque phase a une paire clé/valeur en entrée et en sortie. Les règles de traitement des fonctions "Map" et "Reduce" sont définies par le programmeur .

Hadoop exploite les données en divisant les données d'origine en morceaux de taille fixe appelés fractionnements d'entrée. Un mappeur est défini pour chaque fractionnement et une paire clé/valeur sont générée par une fonction "Map". Les clés et les valeurs sont triées et sont fusionnées lorsque la clé est la même. Ensuite, un combinateur additionne les valeurs de toutes les clés uniques pour chaque mappeur distribué. Cette sortie représente la paire clé/valeur intermédiaire [22].

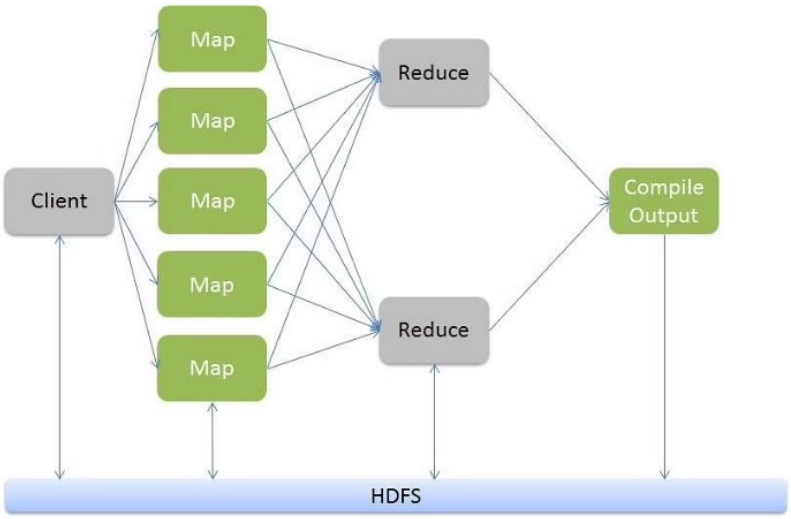

FIGURE 1.3 – Architecture du MapReduce [3]

1.8.2 Apache Spark

Apache Spark est un Framework Open Source [23] de calcul distribué des données volumineuses conçu pour un traitement analytique rapide et sophistiqué. Spark permet d'écrire rapidement des applications en langages Java, Scala ou Python.Il prend en charge les requêtes SQL, les données en continu, l'apprentissage automatique et le traitement des données graphiques.Il est très rapide en mémoire et sur disque par rapport à "MapReduce". Le flux de travail est optimisé par un moteur de Graphe Acyclique Dirigé (DAG).Spark peut-être utilisé avec Elastic MapReduce, HDFS (Hadoop Distributed file System), S3, H. Base, Sequoia DB, MongoDB et d'autres bases de données [23].
Spark se compose du programme pilote, le gestionnaire de cluster et les nœuds de travail. Le principal avantage est qu'il offre un soutien pour déployer des applications Spark dans un cluster Hadoop existant [24].

1.8.3 Apache Storm

Storm est un système de calcul en temps réel distribué et tolérant aux pannes pour le traitement de données volumineuses en continu [25]. Il est spécialement conçu pour le traitement en temps réel contrairement à Hadoop qui est utilisé pour le traitement par lots. De plus, il est facile à configurer et à utiliser, évolutif, tolérant aux pannes pour fournir des performances compétitives.
Sur le cluster Storm, les utilisateurs exécutent les différentes topologies pour différentes tâches alors que la plate-forme Hadoop implémente des travaux de réduction de carte pour les applications correspondantes. Il y a des différences entre les travaux de réduction de carte et les topologies. La différence fondamentale est que le travail de réduction de carte finit par se terminer alors qu'une topologie traite les messages tout le temps, ou jusqu'à ce que l'utilisateur y mette fin.

1.9 Recherches ouvertes dans le Big Data

1.9.1 Internet des Objets

L'Internet a restructuré les interrelations mondiales. En effet, des machines entrent en scène pour contrôler d'innombrables appareil et créent ainsi une nouvelle technique " l'Internet des objets (IoT)". Ainsi, les appareils deviennent les utilisateurs d'Internet, juste comme les humains avec les navigateurs Web. Récemment, (IoT) a attiré l'attention des chercheurs pour ses aspects et ses défis prometteurs. Cette technologie a un impact économique et social impératif pour l'avenir des technologies de l'information, des réseaux et de la communication [24].

1.9.2 Cloud Computing

Le Cloud Computing [20] également appelé informatique en nuage, est un modèle de prestation de services informatiques qui permet l'accès à des ressources informatiques partagées via internet. Plutôt que d'héberger et de gérer des infrastructures informatiques locales, le Cloud Computing permet aux utilisateurs d'accéder à des ressources telles que des serveurs, des bases de données, du stockage, des logiciels et des services via des fournisseurs de services Cloud. Le développement du Cloud Computing est étroitement lié à celui du Big Data. Des architectures plus agiles et plus puissantes sont requises pour optimiser les ressources et assurer la capacité des infrastructures à tenir la montée en charge sans faire exploser les dépenses d'investissement et de maintenance. L'hébergement et les opérations peuvent être externalisés avec le Cloud [20].

1.9.3 Informatique Bio-Inspirée et Informatique Quantique

L'informatique bio inspirée est une technique inspirée de la nature pour résoudre des problèmes complexes du monde réel Biologique. Les systèmes sont auto-organisés sans contrôle central. Un mécanisme de minimisation des coûts bio-inspiré cherche et trouve la solution de service de données optimale compte tenu du coût de la gestion des données et de la maintenance du service. L'informatique bio-inspirées peuvent contribuer à la faisabilité des modèles traditionnels d'exploration de données pour les grands ensembles de

données sous diverses stratégies [26] .
D'autre part l'informatique quantique est comme l'informatique classique, stockant les données sous forme de bits, c'est-à-dire 0 ou 1. Cependant, elle encode les informations dans le bit quantique ou le qubit, ce qui en fait possible de stocker plusieurs états de données simultanément. Plusieurs possibilités s'offrent au domaine de l'intelligence Artificielle avec l'informatique quantique, comme le traitement d'énormes ensembles de données complexes et l'évolution d'algorithmes pour un meilleur apprentissage, raisonnement et compréhension [27].

1.9.4 Analyse de Big Data

L'analyse de Big Data fait référence à l'utilisation de techniques et d'outils d'analyse pour extraire des informations précieuses à partir de grands ensembles de données complexes, variées et volumineuses, connus sous le nom de Big Data. L'objectif principal de l'analyse de Big Data est de découvrir des modèles, des tendances et des informations cachées qui peuvent être utilisés pour prendre des décisions éclairées, améliorer les performances et obtenir un avantage concurrentiel.

1.10 Conclusion

Le « Big Data Médical» peut être utilisé pour l'identification de facteurs de risque de maladies, aider au diagnostic de cancers, au choix et au suivi de l'efficacité des traitements, surveillance des agents pathogènes et prévenir la propagation de maladies mortelles. C'est une source fantastique qui permet de nouvelles connaissances et de progrès médicaux pour l'amélioration de la vie humaine.
Le domaine de l'informatique de la santé bénéficie du développement rapide d'outils d'analyse de données massives pour résoudre des problèmes critiques. Une attention particulière est portée à la gestion des données médicales afin d'assurer leur qualité et leur fiabilité, paramètres cruciaux pour la prise de décision clinique, la sécurité des patients, la recherche médicale, l'évaluation de la performance et de la qualité des soins. Cependant, un défi majeur de l'analyse des données et des problèmes de missing data et représente la problématique traitée dans ce mémoire.

Chapitre 2

SYNTHESE DES TRAVAUX

2.1 Introduction

Ce chapitre va présenter les nouvelles recherches pour la problématique dans le domaine "Données manquantes dans le Big Data médical" de l'année 2014 à 2022, collectées en utilisant le moteur de recherche académique "Google Scholar". Âpres avoir exposer les modèles de données manquantes, nous discuterons des méthodes analytiques prédictives de l'intelligence artificielle pour surmonter les difficultés du manque de données dans le Big Data médical et pour une prise de décision efficace.

2.2 Modèle de données manquantes

Les modèles de données manquantes décrivent les valeurs manquantes et observées dans un ensemble de données. Cependant, il n'y a pas de liste standard des modèles de données manquantes. Dans cette sous-section, nous discutons de trois modèles.

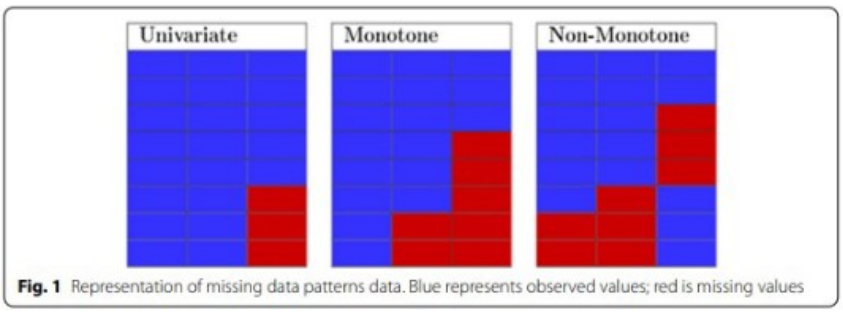

FIGURE 2.1 – Volume annuel global des données [4]

2.2.1 Univarié

Le modèle de données manquantes est univarié lorsqu'il n'y a qu'une seule variable avec des données manquantes. Ce modèle est rare dans la plupart des disciplines et apparaît dans les études expérimentales [4]. Ainsi pour une variable Y_k seulement, si une observation Y_{ki} est manquante, alors il n'y aura plus d'observations de cette variable.

2.2.2 Monotone

On dit que le modèle de données manquantes est monotone si les variables des données peuvent être organisées. Ce modèle est habituellement associé à des études longitudinales où les membres abandonnent et ne reviennent jamais. Un ensemble de données aurait un modèle manquant monotone quand il est possible d'organiser les variables de sorte que, si une personne n'a pas de variable dans ce cas, toutes les variables suivantes sont également manquantes [28]. Le modèle de données monotones est plus facile à traiter puisque les valeurs manquantes sont facilement observables.

2.2.3 Non monotone

Les valeurs manquantes sont non monotones (ou arbitraires) dans le cas d'un modèle dans lequel l'absence d'une variable n'affecte pas l'absence d'autres variables [28]. Dans ce cas, on définit la matrice des valeurs manquantes par M = (m_{ij}) avec m_{ij}= 1 si y_{ij} est manquant sinon zéro.

2.3 Approches pour le traitement des données médicales manquantes

2.3.1 Méthodes de suppression des données (Deletion)

L'approche la plus courante à l'égard des données manquantes consiste à omettre les cas comportant des données manquantes et à analyser les données restantes [29], connues sous le nom de suppression par liste. Une autre approche existante connue sous le nom de

suppression par paires où seules les observations manquantes sont ignorées[29], et l'analyse est effectuée sur les variables présentes. D'autre part, s'il manque trop de données pour une variable, il peut être possible de supprimer la variable ou la colonne de l'ensemble de données. Une analyse appropriée des données est nécessaire avant que la variable ne soit complètement supprimée (vérifier la performance du modèle).

2.3.2 Méthodes d'imputation simple

2.3.2.1 Moyenne, Médiane, Mode, Valeurs Arbitraire

Dans cette technique d'imputation, l'objectif est de remplacer les données manquantes par des estimations statistiques des valeurs manquantes.

A. Moyenne : Dans une substitution moyenne, la valeur moyenne d'une variable est utilisée à la place de la valeur de données manquantes pour cette même variable [30]. L'utilisation de la moyenne est une estimation raisonnable pour sélectionner une observation aléatoire à partir d'une distribution normale [28]. Certains inconvénients de l'imputation moyenne est une surestimation de la taille de l'échantillon, une sous-estimation de la variance, et la corrélation pourrait être négativement biaisée [31].
B. Médiane : Peut-être utilisée lorsque la variable a une distribution asymétrique [4].
C.Mode : L'intérêt d'estimer le mode est de remplacer la population de valeurs manquantes par la valeur la plus fréquente, car c'est l'occurrence la plus probable [30].
D.La valeurs arbitraire : L'imputation arbitraire des valeurs manquantes qui consiste à remplacer toutes les occurrences de valeurs manquantes dans une variable par une valeur arbitraire (différente de la médiane /mode). Les valeurs arbitraires couramment utilisées sont 0, 999, - 999 (ou d'autres combinaisons de 9) [32].

2.3.2.2 Méthodes d'imputation Variée

Il existe d'autres méthodes d'imputation très connues qui sont toutes dans la même thématique telle que :
A.L'imputation par point commun : sachant que pour une échelle de notation, on utilise le point médian ou la valeur la plus couramment choisie [33]. Semblable à l'imputation

par la valeur moyenne, cette méthode est plus appropriée pour les valeurs ordinales.

B.L'imputation par ajout d'une catégorie : elle est beaucoup plus adaptée aux variables catégorielles, ainsi il est possible de traiter le problème des données manquantes très simplement en ajoutant une catégorie supplémentaire pour "manquant" [34].

C.L'imputation par catégorie fréquente : c'est le remplacement des valeurs manquantes par la catégorie la plus fréquente (équivaut à une imputation moyenne /médiane). Elle consiste à remplacer toutes les occurrences de valeurs manquantes dans une variable par l'étiquette ou la catégorie la plus fréquente de la variable [35].

D.La dernière observation reportée : elle est utilisée si les données sont des séries chronologiques, la valeur manquante est remplacée par la dernière valeur observée [36].

E.Interpolation Linéaire : qui remplace les valeurs manquantes par une interpolation linéaire. La dernière valeur valide avant la valeur manquante et la première valeur valide après la valeur manquante sont utilisées pour l'interpolation [37].

F.Imputation Par échantillonnage aléatoire : est semblable à l'imputation moyenne/médiane parce qu'elle vise à préserver les paramètres statistiques de la variable originale, pour laquelle des données sont manquantes, est de prendre une observation aléatoire à partir des observations disponibles et d'utiliser cette valeur extraite au hasard pour combler l'écart [38].

2.3.3 Imputation multiple

La méthode d'imputation multiple (IM) produit une gamme de m valeurs possibles, (m > 1), appelées données imputées, à partir des données existantes. Elles sont analysées à l'aide de certaines méthodes statistiques standard pour obtenir l'ensemble le plus approprié des valeurs des données manquantes. La méthode (IM) aide à restaurer la variabilité naturelle des valeurs manquantes, produit une inférence statistique valide et génère des résultats appropriés en présence d'un volume élevé de valeurs manquantes [31].

2.3.4 Imputation basée sur l'apprentissage automatique (Machine learning)

L'Apprentissage automatique (Machine Learning) ou apprentissage automatique est une sous-catégorie de l'Intelligence Artificielle (IA). Elle consiste à laisser des algorithmes découvrir des patterns , à savoir des motifs récurrents dans les ensembles de données.On énumère ci-dessous les modèles les plus utilisés dans le domaine médical. On a maintenue la dénomination "anglaise ", vue qu'elle est souvent utilisée dans la littérature scientifique.

2.3.4.1 K-Nearest Neighbor imputation (KNN)

K-NN proposé par Cover T et Hart P en 1967, impute les valeurs d'attributs manquantes en fonction du nombre de voisins K les plus proches. Les voisins sont déterminés selon une mesure de distance. Une fois que K voisins sont déterminés, la valeur manquante est imputée en prenant la moyenne/médiane ou le mode des valeurs d'attributs connues [39].

2.3.4.2 Random Forest imputation (RF)

Random Forest [40] est une méthode d'imputation non paramétrique applicable à divers types de variables qui a également été appliquée aux données manquantes. Cette méthode utilise plusieurs arbres de décision pour estimer les valeurs manquantes. Random Forest fonctionne mieux avec de grands ensembles de données car elle peut entraîner un ajustement excessif sur de petits ensembles [40] .

2.3.4.3 Linear Regression imputation (LR)

Dans l'imputation par régression, les variables existantes sont utilisées pour prédire la valeur manquantes [41] [42]. Cette approche présente plusieurs avantages parce qu'elle conserve une grande quantité de données comparées à la liste ou la suppression par paires et évite des changements importants dans l'écart-type (c.-à-d. la dispersion des données autour de la moyenne). Toutefois, comme dans la substitution moyenne, aucune nouvelle information n'est ajoutée, alors que la taille de l'échantillon peut être considérablement augmentée et que l'erreur type (lmesure de l'erreur d'estimation) est réduite.

2.3.4.4 Fuzzy Approches imputation (FCM)

Fuzzy K-Means (FKM) [43] est une méthode utilisée pour remplacer les valeurs manquantes dans un ensemble de données en utilisant des techniques de logique floue. C'est un algorithme de clustering ou chaque point de données peut appartenir à plusieurs clusters avec différents degrés d'appartenance. L'algorithme FKM fonctionne en assignant itérativement les points de données aux clusters et en ajustant les centres de clusters en fonction du degré d'appartenance. Le but d'utiliser FCM au lieu d'un autre algorithme de clustering est en raison de la capacité de cloisonner ou de regrouper les données ambiguës en ayant la valeurs de la fonction d'adhésion [44].

2.3.4.5 Decision Trees imputation (DT)

Decision Trees imputation est un apprentissage supervisé populaire. L'algorithme est utilisé pour construire des modèles de classification et de régression qui divise l'ensemble de données en sous-ensembles (appelés nœuds de décision). Les nœuds qui ne peuvent pas être séparés sont appelés nœuds terminaux ou des feuilles. La méthode utilise un arbre de décision pour prédire les valeurs manquantes dans un ensemble de données. Les variables avec des valeurs manquantes sont utilisées comme cibles, et l'arbre est construit en utilisant les autres variables disponibles. Une fois l'arbre construit, il est utilisé pour prédire les valeurs manquantes en parcourant l'arbre jusqu'à atteindre une feuille et en utilisant la valeur de la feuille comme "estimation". Ce processus est répété pour chaque variable avec les valeurs manquantes [31] .

2.3.4.6 Support Vector Methods imputation (SVM)

Les méthodes de Support Vector Machines (SVM) [45] permettent de résoudre des problèmes de classification ou de régression en construisant un modèle mathématique basé sur des données d'entraînement. L'objectif principal de SVM est de trouver la meilleure séparation possible entre les différentes classes de données, en utilisant des vecteurs de support qui définissent la frontière de décision optimale. En d'autres termes, SVM cherche à trouver un hyperplan qui maximise la marge de séparation entre les différentes classes, permettant ainsi de classifier de nouvelles données. Cet hyperplan est construit en utilisant un sous-ensemble de données d'entraînement, appelées vecteurs de support, qui sont

les points les plus proches de la frontière de décision [45].

2.3.4.7 Maximum Likehood (ML)

Maximum Likehood est une méthode utilisée pour estimer les paramètres d'un modèle à partir des données observées, son objectif est de trouver les valeurs des paramètres qui rendent les données observées les plus "probables" selon le modèle spécifié. Pour appliquer le Maximum Likehood, on suppose généralement que les données suivent une distribution de probabilité spécifique, avec certains paramètres inconnus. L'idée est de trouver les valeurs de ces paramètres qui maximisent la probabilité (ou la vraisemblance) d'observer les données que l'on possède [46].

2.3.4.8 Expectation Maximization (EM)

L'Expectation-Maximisation (EM) est un algorithme itératif utilisé pour estimer les paramètres d'un modèle statistique lorsque les données contiennent des variables latentes non observées. L'algorithme EM est particulièrement utile lorsque certaines données sont manquantes ou lorsque certaines variables ne peuvent pas être directement observées. L'idée principale de l'algorithme EM est de trouver une estimation des paramètres en alternant entre deux étapes : l'étape d'espérance (Expectation) et l'étape de maximisation (Maximisation) [47]. L'étape d'espérance consiste à estimer les valeurs des variables latentes (non observées) à partir des paramètres actuels du modèle. L'étape de maximisation consiste à ajuster les paramètres du modèle en maximisant la fonction de vraisemblance conditionnelle des données observées, compte tenu des valeurs estimées des variables latentes obtenues lors de l'étape précédente. En itérant entre ces deux étapes, l'algorithme EM cherche à trouver les valeurs des paramètres qui maximisent la vraisemblance des données observées, même en présence de variables latentes non observées [47].

2.3.4.9 Apprentissage profond (Deep learning)

L'apprentissage profond ou Deep learning est dérivé de l'apprentissage automatique (machine learning) où la machine est capable d'apprendre par elle-même. Les modèles d'Apprentissage profond utilisent des réseaux neuronaux comportant un grand nombre de

couches et utilisent à la fois des fonctions d'activation et des fonctions de perte dans leur processus d'apprentissage.

Fonctions d'activation Les fonctions d'activation [48, 49] sont utilisées pour introduire de la non-linéarité dans le modèle, ce qui permet aux réseaux de neurones d'apprendre et de modéliser des relations complexes et non linéaires entre les entrées et les sorties. Il existe plusieurs types de fonctions d'activation couramment utilisées, notamment la Sigmoïde, la Tangente Hyperbolique (tanh), ReLU, Leaky ReLU, etc. Chaque fonction d'activation a ses propres caractéristiques et est adaptée à des scénarios spécifiques. Le choix de la fonction d'activation dépend du problème à résoudre et des propriétés souhaitées du modèle.

Fonctions de perte Les fonctions de perte [50, 51] servent à mesurer l'écart entre les prédictions du réseau et les valeurs réelles correspondantes. Elles fournissent une mesure quantitative de l'erreur commise par le réseau sur un exemple d'entraînement spécifique. L'objectif est de minimiser cette fonction de perte afin d'ajuster les paramètres du modèle et d'améliorer progressivement les performances du réseau. Il existe plusieurs fonctions de perte couramment utilisées dans les réseaux de neurones, et le choix de la fonction de perte dépend du type de problème à résoudre. Voici quelques-unes des fonctions de perte les plus utilisées :Erreur Quadratique Moyenne (MSE) ,Erreur Absolue Moyenne (MAE),Entropie croisée binaire, etc.

Les sections suivantes explorent des typologies de réseau neuronal artificiel les plus populaires. On énumère ci-dessous les modèles les plus utilisés pour l'imputation des données manquantes dans le domaine médical :

1- Self-Organizing Map imputation (SOM) L'imputation Self-Organizing Map (SOM) est une méthode utilisée pour remplir les valeurs manquantes dans un ensemble de données. Elle repose sur l'utilisation d'une carte auto-organisatrice, un type de réseau neuronal qui organise et représente les données multidimensionnelles. L'algorithme SOM entraîne la carte en utilisant les données complètes, y compris les valeurs manquantes, pour capturer la structure sous-jacente. Ensuite, les valeurs manquantes sont estimées en se basant sur les valeurs des nœuds voisins sur la carte. Cela permet de compléter les

données manquantes de manière efficace et automatique [52].

2- Multi-Layer Perceptron (MLP) Le Perceptron Multi-Couche (MLP) est un type de réseau neuronal artificiel composé de plusieurs couches de neurones. Il est utilisé pour effectuer des tâches d'apprentissage supervisé telles que la classification et la régression [53]. Le MLP est caractérisé par une couche d'entrée qui reçoit les données, une ou plusieurs couches cachées qui effectuent des calculs intermédiaires, et une couche de sortie qui produit les prédictions. Chaque neurone dans le MLP est connecté aux neurones des couches adjacentes par des poids ajustables. L'apprentissage du MLP se fait par rétropropagation du gradient, où les poids sont ajustés pour minimiser l'erreur entre les prédictions du réseau et les valeurs cibles. Le MLP est capable d'apprendre des relations non linéaires complexes et est largement utilisé dans des domaines tels que la vision par ordinateur, le traitement du langage naturel et la reconnaissance de formes [53].

3- Convolutional Neural Network (CNN) Un réseau de neurones convolutif (CNN) [43] est un type de réseau neuronal spécialisé dans le traitement de données structurées en grille, comme des images. Il utilise des couches de convolution pour détecter des motifs et des caractéristiques locales, et des couches de "pooling" pour réduire la dimension des données. Les CNN sont largement utilisés pour des tâches de vision par ordinateur, telles que la classification d'images et la détection d'objets. Ils sont efficaces pour apprendre des représentations hiérarchiques et capturer des motifs complexes dans les données.

4- Generative Adversarial Networks (GAN) Parmi les méthodes d'imputation RNA figurent ceux qui s'appuient sur les réseaux adversatifs génératifs (GAN), qui produisent de fausses données réalistes par la formation adversielle, c'est l'un des modèles les plus appréciés. Les réseaux antagonistes génératifs (GAN) permettent d'apprendre des représentations profondes sans données de formation. Les GANs[54, 55, 56, 57]sont basés sur deux réseaux de neurones artificiels qui s'affrontent, un générateur et un discriminateur. Ces deux réseaux permettant de générer de l'information : un générateur qui produit des données synthétiques et un discriminateur qui essaie de distinguer les données synthétiques des données réelles. Le but du processus d'entraînement est d'améliorer le générateur pour qu'il produise des données synthétiques de plus en plus réalistes (fiables).

La technique pour atteindre ce but est de jouer un jeu minimax à deux joueurs entre le générateur et discrimination sous la contrainte que le générateur tente de confondre le contenu généré et le discrimination tente de distinguer les données réelles de ce qui le générateur crée.

2.4 Table de comparaison et Discussion

Âpres avoir exposer les différentes méthodes pour le traitement des données médicales manquantes, nous avon établie des Tables de comparaison pour chaque approche en soulignant les avantages et inconvénients de chacune. D'après les travaux scientifiques (35) vus précédemment et les articles (Review) suivants [58, 59, 60, 4], une étude comparative des différentes méthodes d'imputation et de leurs avantages et inconvénients a été établie dans les tableaux ci-dessous :

2.4.1 Table de comparaison des méthodes d'imputation statistique

Méthode	Auteur	Avantages	Inconvénients
Suppression Par Listes (List-wise Deletion)	Strobl, E.,et al (2018)	-Une stratégie raisonnable s'il y a plus de 95% de valeur manquantes	-N'est pas une stratégie adapté pour les datasets médicales, il faut éviter de supprimer les données .
Suppression par paire (Pairwise Deletion)	Weaver, B.,et al (2014)		
Moyenne	P, Kumar et. al (2021)	- Simple et facile à réaliser -Ne modifier pas la moyenne de la variable dans l'échantillon.	-Peut conduire a une incohérence. -La distortion de la variance d'origine. -L adistortion de la covariance avec les variables restans
Médiane	Madhu, G.et al. (2020)	-Simple et facile à réaliser	
Mode	Rani,P et al (2021)		
Dernière observation reportée (Last Observation Carried Forward	Yongqiang Tang (2018)	-Facile à comprendre et à communiquer. -Simple à réaliser. -Suppose fortement que la valeur du résultat reste inchangée par les données manquantes.	- la méthode tend à donner des estimations biaisées en raison d'hypothèses irréalistes au sujet des données manquantes.
Interpolation Linéaire	Daberdaku,S.et al. (2020)	-Donne des bons résultats avec des modèles qui ont des données noncomplexes.	Pourrait ne pas fonctionner dans un modèle assez complexe.
Imputation par point commun	Choi, J.et al. (2018)	-Simple et facile à réaliser -Plus approprié pour les valeurs ordinales.	-La distorsion de la variance d'origine. -La distorsion de la covariance avec les variables restantes dans l'ensemble de données.
Imputation par Catégorie Fréquente	Kunzmann, K., et al (2021)	-Facile à communiquer et simple à réaliser.	

TABLE 2.1 – Table de comparaison des méthodes d'imputation statistique

Imputation Par Valeurs Arbitraires	Zhang, Z. (2016)	-Fonctionne raisonnablement bien pour les caractéristiques numériques principalement positives en valeur et pour les modèles arborescents.	-Peut affecter la performance d'un modèle si les données manquantes sont nombreuses
Imputation Par Échantillonnage aléatoire	Giganti, M. J.et al. (2020)	-Préserve les paramètres statistiques de la variable d'origine	-Peut conduire à des incohérences
Imputation multiple	Bartlett, J. W.et al (2020)	-Très flexible et peut gérer des données manquantes de différents types. -Crée des Datasets complètes.	-Difficile à communiquer. -Prend les imputations incertaines en compte.

TABLE 2.2 – Table de comparaison des méthodes d'imputation statistique

2.4.2 Table de comparaison des méthodes basée apprentissage automatique(Machine learning)

Méthode	Auteur	Avantages	Inconvénients
K-Nearest Neighbor imputation (KNN)	Daberdaku,S.et al. (2020)	-Peut prédire les attributs qualitatifs et les attributs quantitatifs -Il n'est pas nécessaire de créer un modèle prédictif pour chaque attribut avec des données manquantes.	-La performance est affectée quand la valeur de 'k' est grande et pour les grands ensembles des données.
Random Forest imptation (RF)	Yang, B.et al (2018)	-Fonctionne mieux avec les grands ensembles de données.	-L'utilisation sur les petits ensembles de données donne un sur-ajustement
Régression Linéaire imputation (RL)	Fedushko, S.,et al (2019)	-Conserve une grande quantité de données. - Evite de modifier considérablement l'écart-type ou la forme de la distribution.	-La véritable distribution du prédicteur est généralement inconnue et nécessite des hypothèses.
Fuzzy Approches imputation (FCM)	Himansu Das , Bighnaraj Naik , Behera , Shalini Jaiswal , Priyanka Mahato , Minakhi Rout c (2020)	- Grâce à l'apport du flou, l'appartenance d'un point de données à un cluster spécifique est donnée par la valeur d'appartenance du point de données à ce cluster. Cela rend la technique efficace.	- nécessité de la connaissance au préalable du nombre de clusters les résultats de FCM -ne semblent pas très stables et cela à cause de la sélection aléatoire des centres

TABLE 2.3 – Table de comparaison des méthodes basée apprentissage automatique

Decision Trees imputation (DT)	SIMON WILLS, CHARLIE J. UNDERWOOD and PAUL M. BARRETT (2020)	- ils peuvent utiliser des données catégoriques et numériques comme variables prédictives - Simplifie les relations complexes entre les variables - il n'est pas nécessaire de les données doivent être conformes à une distribution normale	- Un inconvénient avec la méthode de l'arbre de décision est celle du dépassement des données. - Les méthodes arborescentes sont également sujettes à biais si certaines classes dominent les données
Support Vector Methods imputation (SVM)	Mehrbakhsh Nilashi ,Hossein Ahmadi , Azizah Abdul Manaf, Tarik A. Rashid, Sarminah Samad,Leila Shahmoradi, Nahla Aljojo9 ,Elnaz Akbari (2020)	- svm a une meilleure précision de calcul. - Performant pour la classification . - SVM n'a pas besoin de formation de données complètes dans la construction de la classification Modèles - temps de traitement considérable	- Ne convient pas aux problèmes non linéaires, n'est pas le meilleur choix pour un grand nombre de fonctionnalités.
Maximum Likehood (ML)	Tomita, H.et al (2018)	-Facile à réaliser.	-Un peu compliquer à interpréter.
Expectation Maximization (EM)	Huang, S. F.et al (2020)		-Peut donner des estimations hors rang.

TABLE 2.4 – Table de comparaison des méthodes basée apprentissage automatique

2.4.3 Table de comparaison des méthodes basée apprentissage profond (Deep learning)

Méthode	Auteur	Avantages	Inconvénients
Self-Organizing Map imptation (SOM)	Bain Khusnul Khotimah1, Miswanto and *Herry Suprajitno (2019)	- les algorithmes SOM sont très outils utiles pour l'analyse de données dans différents contextes - SOM est particulièrement bien adapté pour les données de flux	- la flexibilité étant faible presque inexistante au sein de ces architecture
Multi-Layer Perceptron (MLP)	Cheng, C. Y.et al (2020).	-Fournir des prédictions rapides après avoir été entrainé. -Donne des bonnes précisons.	- Lorsque les données manquantes sont plusieurs dans un grand Dataset, de nombreux modèles MLP doivent être construits. - Temps de réponse élevé.
Convolutional Neural Network (CNN)	Daniel Miller, Andrew Ward, Nicholas Bambos David Scheinker (2018)	- il a l'avantage d'être adapté pour la perte de donnée, En outre, la formation du réseau sur une fonction de perte différente, s'adaptera l'accent de compression en conséquence.	- Difficile a développé - formation nécessite généralement beaucoup de données
Generative Adversarial Networks (GAN)	Jinsung Yoon , James Jordon , Mihaela van der Schaar (2018)	- GAN est toujours performant même lorsque le nombre d'échantillons est relativement faible. - GAN est également robuste au nombre de dimensions. - la performance du GAN à mesure que les taux manquants augmentent surpasse systématiquement les indices de référence sur l'ensemble de la gamme des taux manquants. - meilleure précision de prédiction post-imputation	- le temps d'exécution moyen des réseaux génératifs antagoniste est un peu plus long à converger et à finir son apprentissage .

TABLE 2.5 – Table de comparaison des méthodes basée apprentissage profond

2.5 Conclusion

Après la synthèse des travaux dans le domaine (Missing Médical Data) on remarque que les recherches les plus importantes s'insèrent dans laxes des méthodes Basées apprentissage profond (Deep learning) plus précisément les réseaux adversatifs génératifs (GAN) qui présentent plusieurs atouts en matière de précision et de qualité de données imputées. En dépit du fait que le temps d'exécution moyen, ces réseaux GANs [54] est un peu plus long à converger pour un apprentissage idéal. Néanmoins, le type d'exécution représente un obstacle pour ces réseaux et pour le qu'elle on présente une solution dans le chapitre suivante.

Chapitre 3

CONCEPTION ET IMPLÉMENTATION

3.1 Introduction

L'objectif de ce chapitre est de présenter les étapes de conception de notre modèle d'imputation des données manquantes dans les bases de données médicales . Vu l'étude comparative présentée dans le chapitre précédent des différentes méthodes d'imputation de données manquantes et la comparaison établie dans le travail de recherche présentée dans la conférence IAM 2022 [61], on a décidé d'avoir recours au Réseaux Antagonistes Génératifs (GAN) [54]. Dans la section suivante, on expose l'architecture du modèle proposé suivi par l'implémentions et la discutions des résultats obtenus.

3.2 Conception

3.2.1 Architecture proposée

Les Réseaux Antagonistes Génératifs [54] se démarquent des autres méthodes par la qualité et la précision des données imputées (facteur important dans le domaine médical). Sachant que les données médicales sont caractérisées par l'hétérogénéité, on a choisi d'appliquer le modèle Fuzzy K-Means (FKM) [43] pour obtenir de meilleures performances et permettre au modele d'avoir un meilleur apprentissage.

Pour imputer les données manquantes, plusieurs étapes ont été suivies : Tout d'abord le Dataset vas subir un pré-traitement (nettoyage de données). Cette étape est suivie par une classification générée par le modèle Fuzzy K-Means (FKM). Cette méthode permet d'affecter chaque individu avec des données manquantes aux clusters auxquels il peut appartenir. Enfin dans l'étape d'imputation, tous les individus avec des données manquantes sont imputées en utilisant le modèle proposé. L'architecture du modèle proposé ,nommé Fuzzy-GAN(FGAN) est représentée dans la figure suivante puis détaillée dans les sections suivantes.

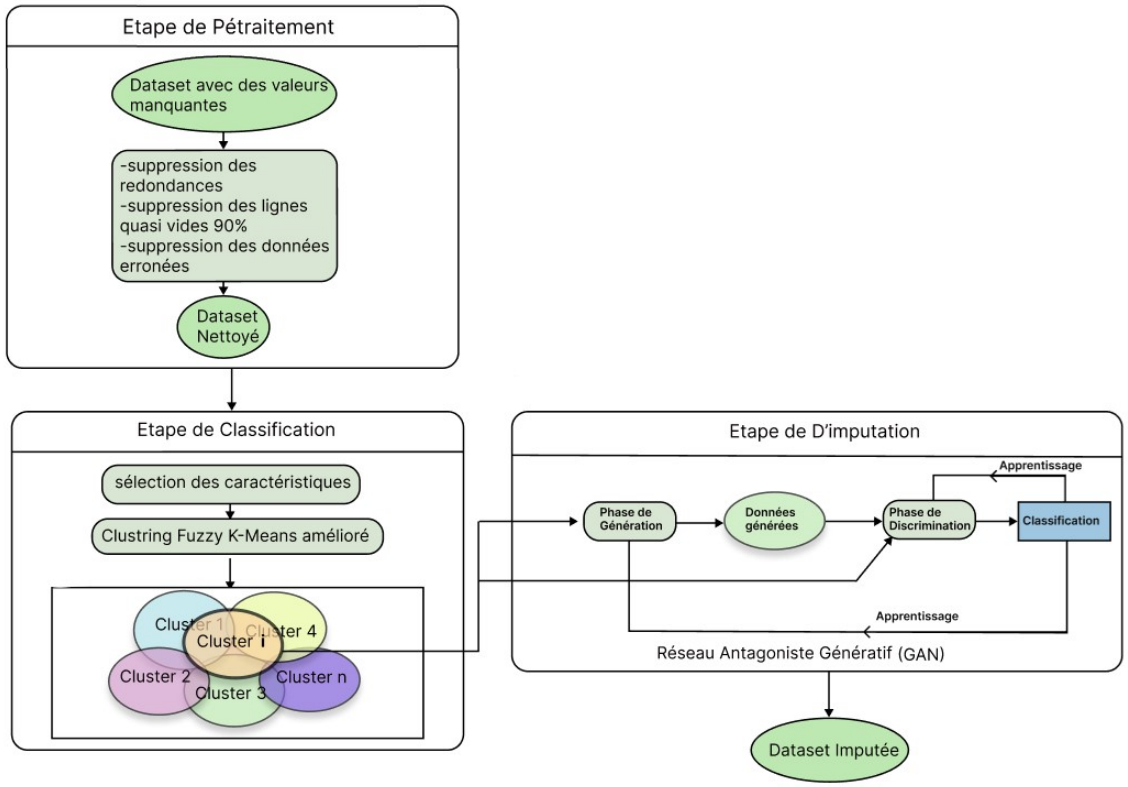

FIGURE 3.1 – Architecture proposée

3.2.2 Étape de prétraitement

Le prétraitement des données (nettoyage de données) est une étape importante qui améliora amplement la qualité des données utilisées dans le modèle. Elle est effectuée comme suit :

Phase 1 : Suppression des lignes qui contiennent beaucoup de données manquantes (90%) on utilisant pour cela la méthode de suppression par listes(list-wise deletion) [29].

Phase 2 : Suppression des redondances (données qui se répètent), en utilisent la méthode prédifinie dans la bibliothéque sklearn en python (sklearn.feature_selection.RFE) pour éliminer les attributs ayant une faible importance prédictive. Cela permet de sélectionner les attributs les plus informatifs et de supprimer les redondances.
Phase 3 : Suppression des données erronées (Valeurs dépassant le seuil reconnu pour les variables), la consultation des références médicales (normes) et experts pour déterminer les seuils acceptables pour chaque variable. Ces seuils peuvent varier en fonction de l'âge, du sexe et d'autres facteurs. Par exemple, pour la glycémie normal d'une personne est de 0,70g jusqu'à 1,10g par litre de sang à jeun.
Cette étape qui a prit beaucoup de temps et de recherches est importante pour obtenir une bonne qualité des données utilisé dans le modèle. Elle est suivie par l'application du modèle de classification proposé qui affectera les individus aux clusters les plus probables.

3.2.3 Étape de classification

Le modèle de classification que nous avons proposé comporte deux étapes : la sélection des caractéristiques (Feature selection) qui permet au modèle de mieux capturer les relations entre les caractéristiques importantes et d'éviter de se concentrer sur des caractéristiques non pertinentes,et le modèle Fuzzy K-Means (FKM) qui utilise une approche floue pour affecter les individus à des clusters en permettant une appartenance partielle à plusieurs clusters .

3.2.3.1 Sélection des caractéristiques (Feature selection)

La sélection des caractéristiques [62] est une technique de réduction de dimensionnalité utilisée pour identifier le nombre optimal de caractéristiques à inclure dans un modèle.

La Variance Inflation Factor (VIF) [63] est une méthode utilisée pour évaluer la multicollinéarité [64] entre les variables d'un modèle. La multicollinéarité se produit lorsque deux ou plusieurs variables indépendantes d'un modèle sont fortement corrélées entre elles.

Cette méthode calcule le rapport de la variance d'une variable dans un modèle incluant toutes les autres variables, par rapport à la variance de cette même variable dans un modèle où elle est seule. Le résultat obtenu permettra de déterminer la dépendance entre

variables. La méthode VIF utilisée sur le dataset lu, calcule la variance selon la formule suivante [63] :

$$V = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \tag{3.1}$$

avec :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.2}$$

Cette étape est suivie par la classification Fuzzy K-Means.

3.2.3.2 Méthode améliorée du Fuzzy K-Means

Fuzzy K-Means (FKM) [43] est un algorithme de clustering ou chaque point de données peut appartenir à plusieurs clusters avec différents degrés d'appartenance. L'algorithme FKM fonctionne en assignant itérativement les points de données aux clusters et en ajustant les centres de clusters en fonction du degré d'appartenance. Le degré d'appartenance est une valeur floue entre 0 et 1, qui indique la force de l'association du point de données avec chaque cluster. L'avantage de L'algorithme FKM par rapport aux K-Means traditionnels [65] est la capacité de gérer les données bruyantes, les grappes qui se chevauchent chose très fréquente dans les datasets.

D'autre part, la faiblesse de l'algorithme Fuzzy K-Means (FKM) est l'initialisation du nomber de cluster « K » qui conditionne le résultat final puisque choisir un nombre de cluster K n'est pas forcément intuitif. Pour y remédier, nous avons utilisé la méthode Elbow [66] pour générer la valeur optimale du nombre des clusters « K », cette méthode améliora amplement la classification. La méthode Elbow utilise la variance au sein de chaque cluster pour définir le nombre K. La variance des clusters se calcule comme suit :

$$V = \sum_{j}\sum_{x_i \to C_i} D\left(c_j, x_i\right)^2 \tag{3.3}$$

Avec :

c_j : Le centre du cluster.

x_i : La ième observation dans le cluster

D(c_j,x_i) : La distance euclidienne entre le centre du cluster c_j et $x_i$le point Après avoir utilisé la méthode Elbow.

Après avoir utilisé la méthode Elbow [66], le nombres de clusters est obtenu et leurs centres

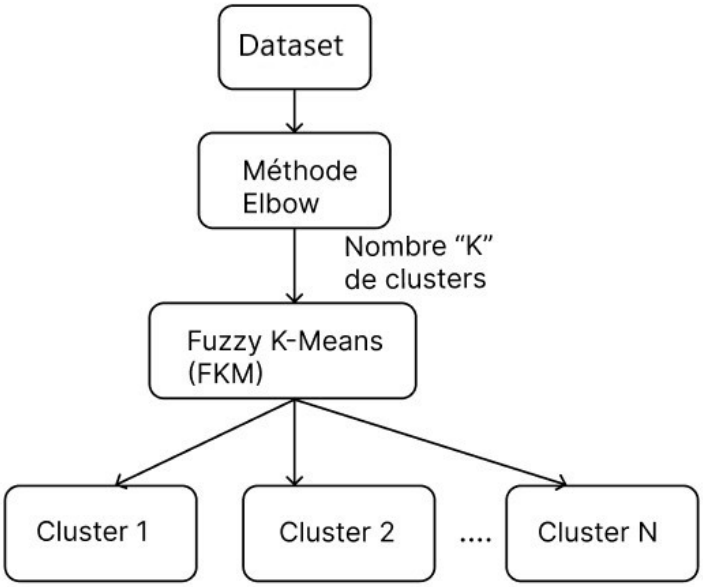

FIGURE 3.2 – Architecture Elbow Fuzzy K-Means

sont calculés. La figure 3.2 montre l'architecture du Fuzzy K-Means (FKM) améliorée par la méthode Elbow.

3.2.4 Étape d'Imputation

Le modèle proposé d'imputation des données manquantes est un réseau antagonistes génératifs communément appeler (GAN). Les GANs [54, 55, 56, 57]sont des modèles de Deep Learning basés sur deux réseaux de neurones artificiels qui s'affrontent, un générateur et un discriminateur. Ces deux réseaux permettant de générer de l'information : un générateur qui produit des données synthétiques et un discriminateur qui essaie de distinguer les données synthétiques des données réelles. Le but du processus d'entraînement est d'améliorer le générateur pour qu'il produise des données synthétiques de plus en plus réalistes et fiables.

On a utilisée la méthode MinMax [67]. Cette dernière est décrite comme un jeu d'optimisation a deux joueurs (dans notre cas le générateur et le discriminateur) où l'objectif est de trouver un équilibre entre les deux. L'objectif est décrit par la génération de données synthétiques aussi semblables que les donnée réelles. Cela signifie que le discriminateur ne peut plus distinguer les données générées des données réelles. L'architecture du modèle proposé, se basant sur les travaux de [55, 56, 57], puis est représentée dans la figure

suivante :

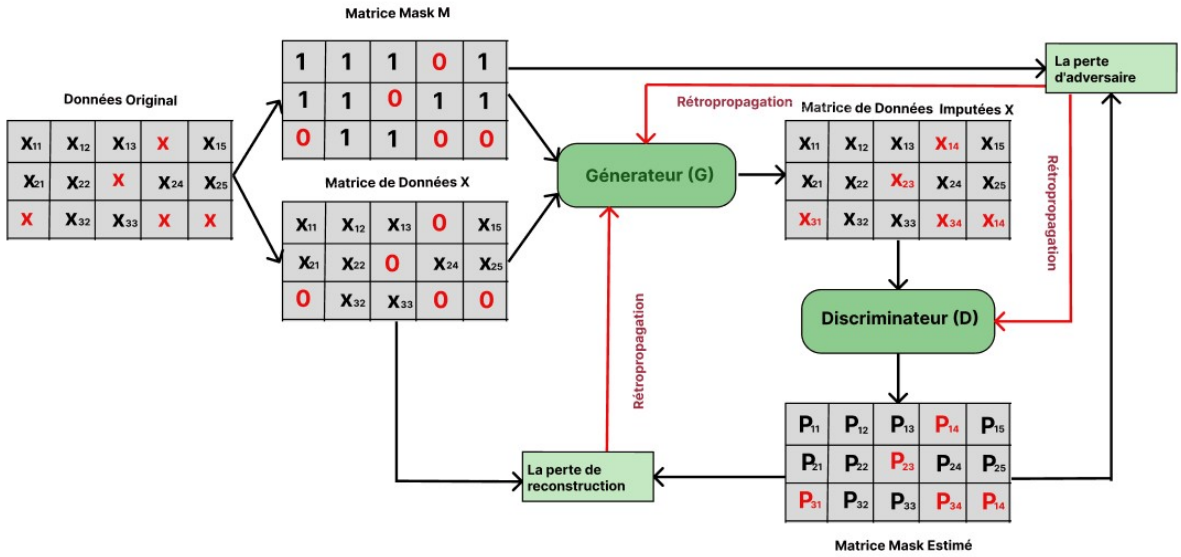

FIGURE 3.3 – Architecture générale du modèle proposée pour l'imputation des données médicales manquantes

3.2.4.1 Fonctionnement général

Pour l'imputation des données médicales manquantes, les données orignales sont représentées par une matrice de données X avec un masque M utilisé pour indiquer les valeurs sont manquantes. Ainsi, [X, M] formera l'entrée du générateur. Après la phase de génération, la matrice de données imputée $\bar{X}$ est obtenue et formera l'entrée du discriminateur. Pendant, la phase de discrimination, le discriminateur va tenter de distinguer les données réelles des données générées et produira une matrice masque $\bar{M}$ qui identifiera les données réelles des données générées selon une estimation définie. Ces deux phase vont être alternées l'une après l'autre jusqu'à que le générateur ait un apprentissage idéal (génération de données parfaites) c.à.d. le discriminateur n'arrive plus à distinguer les données réelles des données générées . Le fonctionnement du générateur et du discriminateur sont détaillés dans les sections suivantes :

3.2.4.2 Générateur (G)

Le Générateur (G) est un réseaux de neurones(MLP) composé de trois couches complètement connectées. La première couche (couche d'entrée) prend en entrée [X,M] et produit

une représentation de dimensions inférieures, qui décrit les données en capturant les caractéristique élémentaires et simple.

La deuxième couche (couche cachée) est responsables de l'apprentissage de représentations et capture les relations complexes et abstraites. Elle permet une compréhension approfondie des données à partir de la représentation den entrée et produit une reconstruction des données manquantes en les projettent dans leur espace d'origine.

Ces couches utilisent des neurones connectés égale au nombre des variables d'entrées et la fonctions d'activation non linéaires Relu [48, 68] qui permet de retourne la valeur d'entrée si elle est positive ou zéro si elle est négative et cela pour capturer des motifs et des structures complexes dans les données générées.

La troisième couche (couche de sortie) du générateur transforme la représentation interne apprise par les couches précédentes en une sortie qui correspond aux caractéristiques et à la structure des données réelles. L'objectif de la troisième couche est de générer des échantillons de données synthétiques qui sont aussi proches que possible des exemples réels.

La qualité de la génération dépend de la capacité du générateur à capturer les caractéristiques et les motifs des données réelles pendant l'entraînement du réseau GAN. Elle utilise la fonction d'activation Sigmoid [48, 51] qui est une fonction non linéaire qui prend une valeur d'entrée x et la transforme en une valeur comprise entre 0 et 1. Cette normalisation des valeurs peut être interprétée comme une probabilité qui facilite la comparaison, l'interprétation et le calcul de mesures d'évaluation telles que l'erreur quadratique moyenne (MSE) utilisée.

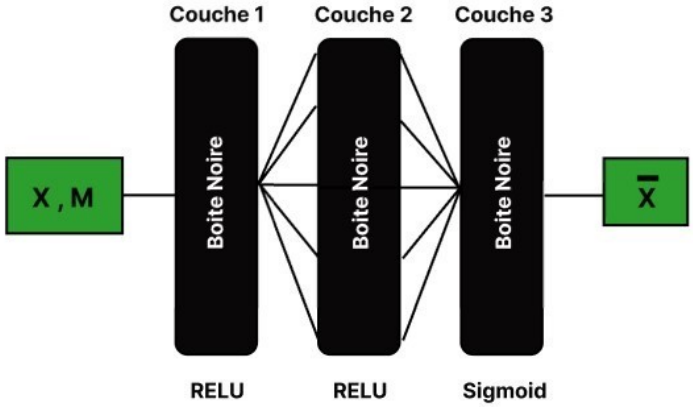

FIGURE 3.4 – Générateur (G)

3.2.4.3 Discriminateur (D)

Le Discriminateur (D) est un réseaux de neurones (MLP) composé de trois couche complètement connecté. La première couche (couche d'entrée) prend comme entrée les données originales et les données terminées générées par le réseau du générateur. L'objectif de cette couche est de fournir une représentation initiale de bas niveau des données en vue de la discrimination en utilisent la réduction de dimensionnalité.
La deuxième couche (couche cachée) du discriminateur prend en entrée les caractéristiques de bas niveau extraites par la première couche et produit avec ces caractéristiques une représentation de haut niveau plus complexe grâce a la reconstruction des données en les projettent dans leur espace d'origine. Ces couches utilisent des neurones connectés égale au nombre des variables d'entrées et une fonctions d'activation non linéaires Relu [48, 68]. Ce choix est justifié dans le paragraphe précédent.
La couche de sortie (couche de sortie) du discriminateur est responsable de la classification finale des données en tant que réelles ou fausses. Elle prend en entrée les caractéristiques de haut niveau obtenues par la deuxième couche et produit une seule valeur de sortie qui représente la probabilité que les données en entrée soient réelles. La fonction d'activation utilisé dans la troisième couches est Sigmoid [48, 51]. Le discriminateur est entraîné pour classer les données complètes et les données incomplètes en utilisant une fonction de perte qui mesure la capacité du discriminateur à distinguer les données complètes des données incomplètes.

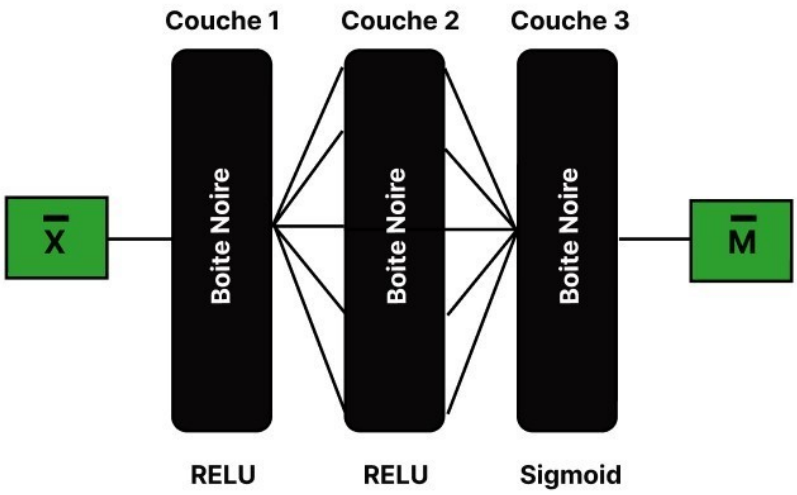

FIGURE 3.5 – Discriminateur (D)

3.2.4.4 Fonction objective du FGAN

La fonction objective [54, 55, 56, 57] du FGAN vise à trouver un équilibre entre le générateur (G) et le discriminateur (D). La fonction objective est définie comme suit :

$$\min_G \max_D V(D, G) \tag{3.4}$$

où V(D, G) est la fonction de perte [54] globale qui mesure la compétition entre le discriminateur et le générateur.La fonction V(D, G) est décomposée en deux termes principaux :perte de reconstruction (reconstruction loss), perte d'adversaire (adversarial loss).

La perte de reconstruction [57] mesure la différence entre les données réelles et les données imputées,en utilisant une fonction de perte comme l'Erreur Quadratique Moyenne (MSE).

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.5}$$

La perte d'adversaire [57]mesure la capacité du générateur à imiter la distribution de données réelle ,alors que pour le discriminateur elle mesure la différence entre l'entropie de la distribution réelle et l'entropie de la distribution imputée. Elle est calculée en utilisant la divergence de Kullback-Leibler (KL) [69] entre la distribution des données réelles et la distribution des données imputées. Elle est présentée par la formule suivante [54] :

$$L_{ADV} = D_{KL}(P_{data}|P_G(Z)) \tag{3.6}$$

P_{data} : la distribution des valeurs complète réelle.
$P_G(Z)$: la distribution des valeurs complète générée à partir d'un vecteur de bruit Z.
D_{KL} : la divergence de Kullback-Leibler entre ces deux distributions.

Le discriminateur D est entraîné pour maximiser la probabilité de prédire correctement les valeurs manquantes M. En d'autres termes, le discriminateur essaie de distinguer les données réelles des données imputées par le générateur. D'autre part, le générateur G est entraîné pour minimiser la probabilité que le discriminateur D prédise les valeurs manquantes M. Le générateur essaie de générer des données qui trompent le discriminateur.

3.3 Implémentation

Dans cette étape, on décrit en premier le matériel, logiciel et DataSet utilisés dans implémentation du modèle proposé. Les résultats obtenus seront aussi présentés et discutés dans les sections suivantes.

3.3.1 Matériels utilisés

L'implémentation de notre système a été réalisée sur une machine possédant les caractéristiques suivantes avec les logiciels présentés ci-dessous :
Processeur : I7
Mémoire : 8,00 Go
Disque dur : 500 GB

3.3.2 Logiciels utilisés

a) Système d'exploitation : Windows 10 Professionnel.
b) Outils de développement :
Python version 3.9.13 : Python est un langage de programmation de haut niveau avec une syntaxe simple et une puissance remarquable.
Les bibliothèques connexes sont :tensorflow, pyspark, pandas, skfuzzy , statsmodels , sklearn , numpy.
c) Apache Spark version 3.1.3 : il permet d'effectuer des traitements sur de large volume de données.

3.3.3 Datasets utilisés

Nous avons utilisé deux datasets médicals pour tester le modèle proposé .Le premier Wisconsin (Diagnostic) qui représente un ensemble de données de «Cancer du Sein » [1].

	Mean Radius	Mean Texture	Mean Perimeter	Mean Area	Mean Smoothness	Mean Compactness	Mean Concavity	Mean Concave Points	Mean Symmetry	Mean Fractal Dimension	...	Worst Radius	Worst Texture	Worst Perimeter	Worst Area	Sm
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	25.380	17.33	184.60	2019.0	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	24.990	23.41	158.80	1956.0	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	23.570	25.53	152.50	1709.0	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	14.910	26.50	98.87	567.7	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	22.540	16.67	152.20	1575.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	25.450	26.40	166.10	2027.0	
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	23.690	38.25	155.00	1731.0	
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	18.980	34.12	126.70	1124.0	
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	25.740	39.42	184.60	1821.0	
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	9.456	30.37	59.16	268.6	

569 rows × 30 columns

FIGURE 3.6 – Dataset Breast

Le deuxième dataset médical est une base de données massive nommé « Pima » qui représente un ensemble de données de « Diabète », provenant du « National Institute of Diabetes and Digestive and Kidney Diseases » [2] Ce Dataset contient des informations sur des femmes issues d'une population proche de Phoenix, en Arizona, aux États-Unis .

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	4	129	70	18	122	29.43	1.17	45
1	1	205	76	36	249	37.28	0.92	29
2	8	97	82	0	0	37.82	0.59	68
3	7	141	90	41	0	34.25	0.40	39
4	4	120	72	0	0	29.12	0.39	46
...	...	...	...	...	...	...	...	...
77563	3	91	58	11	54	26.26	0.27	22
77564	2	112	62	32	56	26.40	0.13	21
77565	4	128	68	0	0	36.47	0.40	29
77566	1	101	68	21	0	28.56	1.11	22
77567	9	169	88	0	0	31.13	0.32	49

77568 rows × 8 columns

FIGURE 3.7 – Dataset Diabètes

1. https ://www.kaggle.com/code/buddhiniw/breast-cancer-prediction/input
2. https ://www.kaggle.com/datasets/pradeepgurav

3.4 Modélisation d'execution avec Platforme Spark

le programme "Pilot" (SparkContext) dans spark [5] est responsable de la création, de la coordination des opérations et de la gestion des ressources au sein d'une application Spark. Il s'agit de l'interface principale entre l'application et le cluster spark, permettant ainsi de tirer parti des fonctionnalités de Spark pour le traitement distribué des données. Le SparkContext est utilisé pour créer les Resilient Distributed Dataset (RDD).
Une RDD est une collection de données immuable et distribuée sur les nœuds d'un cluster spark. Elle permet de traiter des volumes importants de données de manière parallèle et résiliente aux pannes. Après avoir installé tout les logiciels requis, Spark procède à la lecture du dataset en utilisant l'outil RDD tel qu'il est présenté dans la figures suivante :

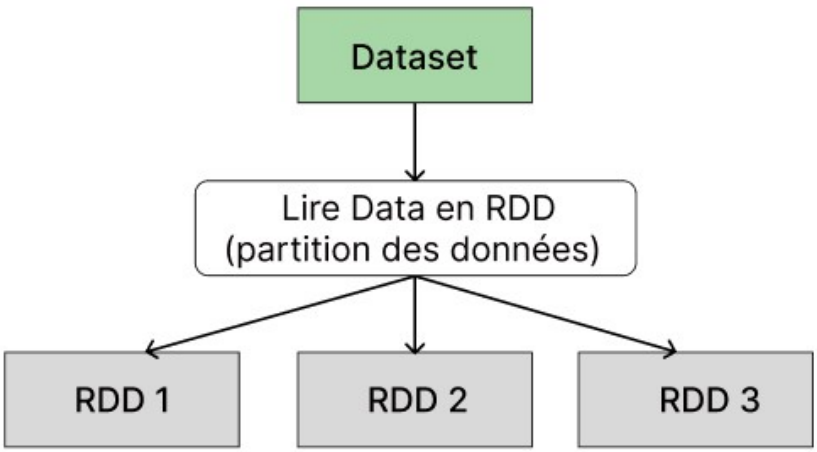

FIGURE 3.8 – Lécture du Dataset en RDD

Un cluster spark[5] est un ensemble de nœuds worker qui peuvent être des machines physiques ou virtuelles. Ils travaillent ensemble pour exécuter un ensemble de tâches sur les données.
Aprés la division des données en partitions RDD, elles sont distribuées sur les nœuds du cluster. Chaque partition est traitée indépendamment et parallèlement par les nœuds worker, permettant ainsi un traitement efficace des données à grande échelle.

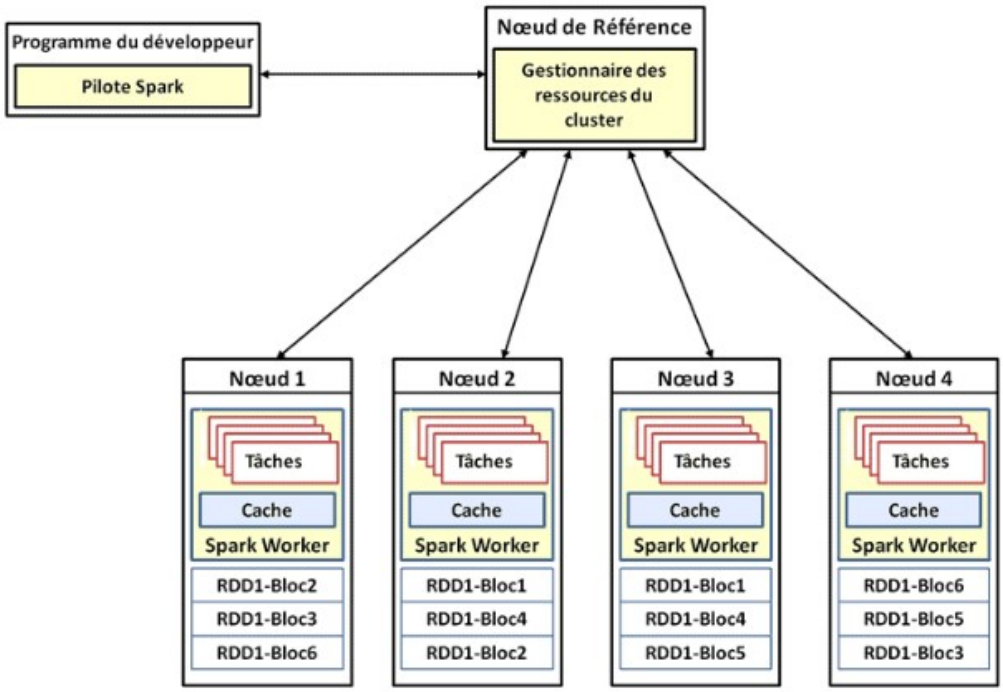

FIGURE 3.9 – Exécution d'une application spark dans un cluster[5].

A la différence d'un cluster Hadoop[20], les processus Workers dans Spark sont des processus d'une durée de vie qui ne s'achève pas à la fin de l'exécution du programme Pilote, ils sont continuellement actifs, c'est pourquoi ils peuvent persister sur les partitions de RDD en mémoire[5].

3.5 Implémentation du modèle proposé

Le modèle proposé a été implémenté et présenté dans toutes ses étapes avec le dataset "Diabètes" ci-dessous :

3.5.1 Sélection des caractéristiques (Feature selection)

Pour la sélection des caractéristiques la méthode « Variance Inflation Factor VIF »[63] est utilisée sur la base de données lu, la figure si-dessous représente les valeur de VIF pour chaque caractéristique :

	variance	Features
0	3.273361	Pregnancies
1	16.547796	Glucose
2	14.597366	BloodPressure
3	4.002236	SkinThickness
4	2.062674	Insulin
5	18.222419	BMI
6	3.192519	DiabetesPedigreeFunction
7	13.474880	Age

TABLE 3.1 – Variances des attribus du Dataset "Diabètes"

En général, les valeurs de VIF [63] supérieures à 5 (valeur empérique par défaut) sont considérées comme indiquant une multicollinéarité importante. Donc cette étape est suivie par la suppression des axes qui ont une variance supérieur a 5.

Unnamed: 0	Pregnancies	SkinThickness	Insulin	DiabetesPedigreeFunction	cluster_nbr
0	4.0	18.0	122.0	1.17	0
1	1.0	36.0	249.0	0.92	1
5	5.0	41.0	42.0	0.16	0
6	1.0	23.0	94.0	0.17	0
12	1.0	14.0	415.0	0.41	3
...	...	...	...	...	...
77555	2.0	15.0	76.0	0.57	0
77557	4.0	12.0	87.0	0.46	0
77560	1.0	46.0	180.0	0.35	2
77563	3.0	11.0	54.0	0.27	0
77564	2.0	32.0	56.0	0.13	0

34037 rows × 5 columns

FIGURE 3.10 – Dataset "Diabètes" après l'utilisation de 'Feature Selection'

Après cette étape il est nécessaire d'appliquer la méthode Elbow[66] pour définir le nombre exacte de cluster pour la division du dataset .

3.5.2 Méthode Elbow

La méthode Elbow[66] a été appliqué sur le data pour générer le nombre de cluster K utile au modèle Fuzzy K-means qui est presenté ci-dessous :

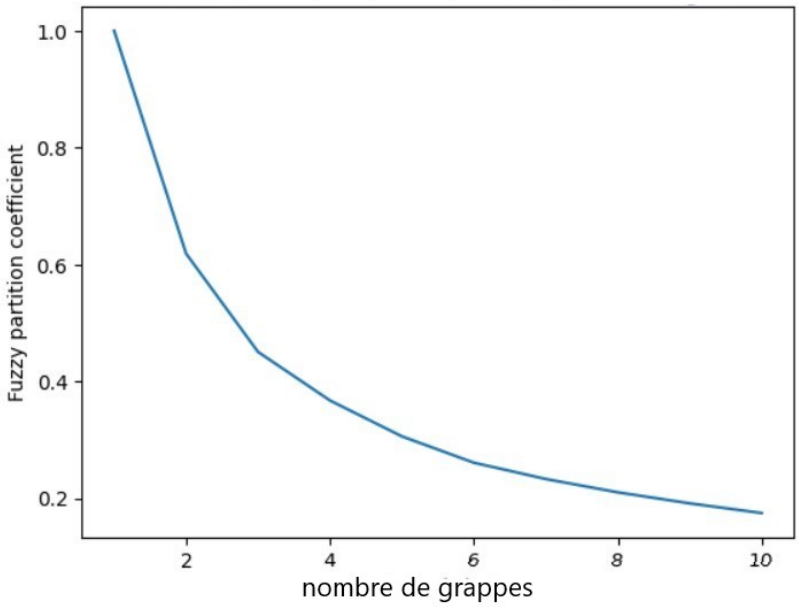

FIGURE 3.11 – Méthode Elbow sur le Dataset "Diabètes"

Comme le montre la figure 3.10 le nombre K est égale à 5 (K=5) , représentant le coude (Elbow)[66] de la courbe obtenue. Le nombre K est utilisée dans la méthode Fuzzy K-means [43] dans l'étape suivante.

3.5.3 Fuzzy K-means

Nous avons regroupé les individus du Dataset avec le Fuzzy K-means en utilisant la valeur du K obtenue précédemment. En utilisant un plan factoriel, on peut projeter les données dans un espace à deux dimensions qui permet de visualiser les relations entre les variables de manière graphique. Cette représentation permet souvent de mieux comprendre la structure des données, d'identifier des groupes ou des motifs dans les données, et de détecter des valeurs aberrantes ou des données manquantes. La figure suivante représente les clusters dans une représentation 2D :

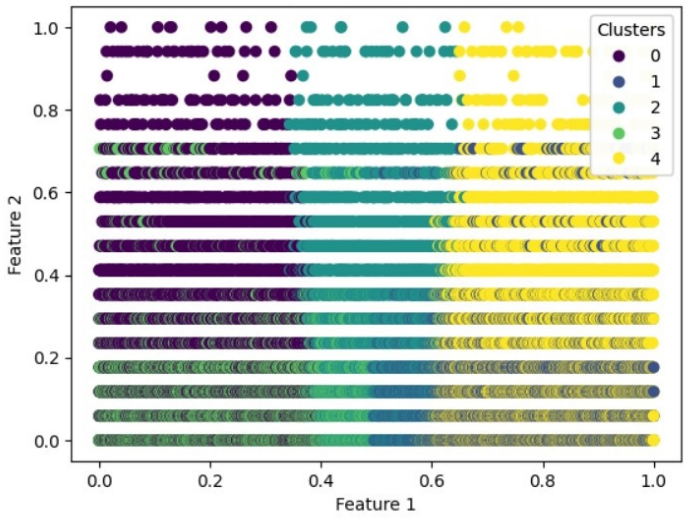

FIGURE 3.12 – Représentation 2D des clusters Fuzzy K-means du Dataset "Diabètes"

Une représentation 3D est utile pour visualiser aussi les données qui ont plus de deux variables, car elle permet de représenter les données dans un espace tridimensionnel. Contrairement à une représentation 2D, une représentation 3D peut montrer des relations complexes entre trois variables ou plus, et permet donc une analyse plus approfondie des données. Voici une représentation 3D pour une meilleur visualisation :

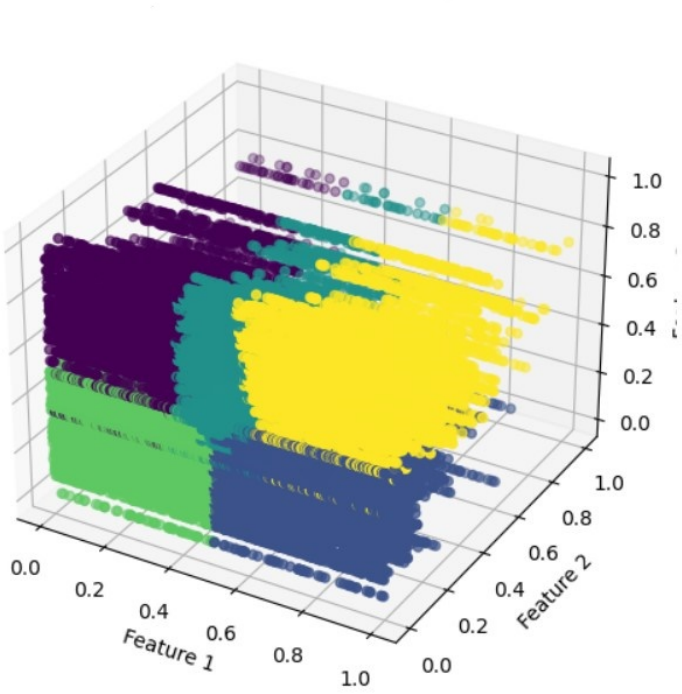

FIGURE 3.13 – Représentation 3D des clusters Fuzzy K-means du Dataset "Diabètes"

3.5.4 Réseau Antagoniste Génératif (GAN)

Le réseau antagoniste génératif (GAN) est utilisé pour l'imputation des données médicales manquantes existantes dans chacun des grappes générés précédemment par la méthodes Fuzzy K-means. En utilisant un générateur et un discriminateur, le GAN peut apprendre la distribution des données existantes et générer de nouvelles données qui ressemblent à celles d'origine.

3.6 Discussion des résultats FGAN

Pour bien analyser les résultats de notre modèle d'imputation proposée (FGAN), il a fallu le tester dans plusieurs circonstances pour souligner sa fiabilité en question de qualité de données imputées et vérifier sa robustesse par rapport a des taux de données manquantes croissant. Ensuite la comparaison du modèle (FGAN) avec différentes méthodes d'imputation, a permis une meilleur compréhension de ses performances et la visualisation de l'amélioration offerte par le modèle proposé.

3.6.1 Imputation par FGAN sur des données connus

Tout d'abord pour vérifier la qualité des données générées par le modèle, nous avons utilisé une colonne du dataset Breast (Mean Concavity) a la qu'elle 20% de données ont été supprimé aléatoirement. Pour dévoiler la précisons du modèle, les données initiales de la colonne on été comparé a l'échantillon généré par notre modèle FGAN. La figure ci-dessous représente la courbe de l'échantillon avant et après imputation par FGAN :

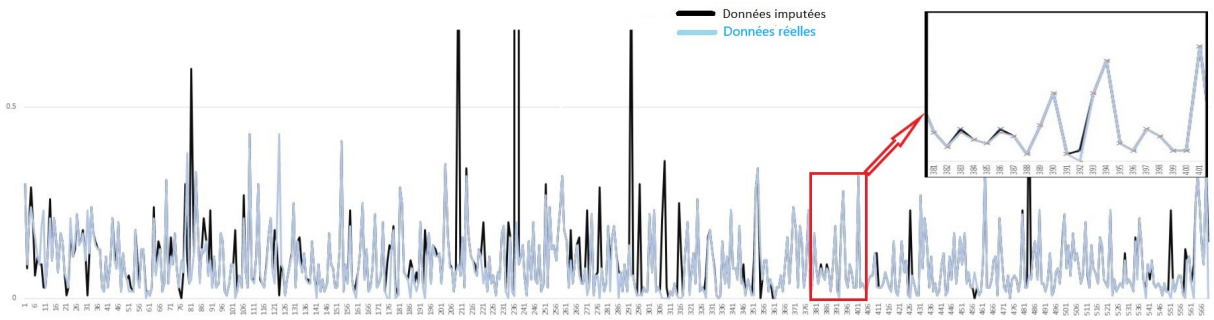

FIGURE 3.14 – Échantillon avant et après imputation par le modèle FGAN

Pour une meilleur compréhension de la partie agrandie dans la figure 3.15 le tableau ci-dessous représente les valeurs des 20 cases (valeurs sélectionnées) :

Numéro de case	Data avec valeur manquantes	Données imputées	Données originales
381	0	0.04	0.04
382	0	0.0852	0.08
383	0.06	0.06	0.06
384	0	0.05	0.05
385	0	0.09	0.08
386	0.07	0.07	0.07
387	0.02	0.02	0.02
388	0.1	0.1	0.1
389	0.19	0.19	0.19
390	0.02	0.02	0.02
391	0	0.033	0
392	0.19	0.19	0.19
393	0.28	0.28	0.28
394	0	0.05	0.05
395	0	0.03	0.03
396	0.09	0.09	0.09
397	0.07	0.07	0.07
398	0.03	0.03	0.03
399	0.03	0.03	0.03
400	0.32	0.3081	0.32
401	0.03	0.03	0.03

TABLE 3.2 – Échantillon de 20 cases

On remarque que les deux courbes sont très similaires sur la majorité des points ce qui reflète que notre modèle génère les données manquantes d'une manière très fiables et très identiques puisque les deux courbes sont très proches l'une de l'autre .
Additionellement la distance euclidienne moyenne[70] est utilisée pour évaluer la qualité des prédictions du modèle proposé par rapport aux véritables valeurs. avec la mesure de la dissimilarité entre les valeurs réelles et prédites.

$$Distance\_euclidienne = \sqrt{\frac{\sum_{i=1}^{n}(B_i - A_i)^2}{n}} \tag{3.7}$$

- A : Données originales
- B : Données imputées
- n : Nombre cases

Avec l'échantillon précèdent on obtient la distance euclidienne égale a (0.001), cette valeur démontre la grande ressemblance entre les valeurs réelles et générées .

3.6.2 Imputation par FGAN sur des taux de données manquantes croissants

Le FGAN apprend à générer les données manquantes à partir des données initiales donc plus il dispose d'informations plus la qualité de données sera grande et plus le taux d'erreur RMSE [71] sera petit. RMSE (Root Mean Square Error) est une mesure couramment utilisée pour évaluer la précision d'un modèle de régression ou de prévision. Il représente l'écart quadratique moyen entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible[71]. Dans la figure ci-dessous on a appliqué le modèle proposé FGAN sur le dataset (Breast) et la base de données massives (Diabètes) avec des taux de données manquantes variant de 20% jusqu'à 80% générée a l'aide de la méthode prédéfinie (Random) dans la bibliothèque Numpy pour choisir des cases aléatoirement puis les remplacé avec zéro :

Breast	20%	30%	40%	50%	60%	70%	80%
RMSE	0.18	0.19	0.20	0.203	0.21	0.23	0.24
Diabtes	20%	30%	40%	50%	60%	70%	80%
RMSE	0.18	0.18	0.19	0.195	0.20	0.215	0.23

TABLE 3.3 – Tableau d'erreur RMSE du modèle FGAN appliqué sur les Dataset "Breast" et "Diabètes" avec des taux de valeurs manquantes croissants

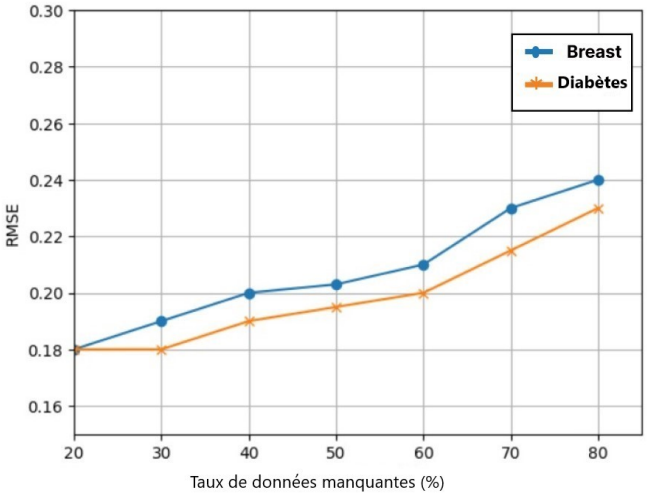

FIGURE 3.15 – Courbes d'erreur RMSE sur les Dataset "Breast" et "Diabètes" avec des taux de valeurs manquantes croissants

Le modèle proposé FGAN donne une valeur initiale excellente de RMSE soit (0.18) pour

les deux datasets contenant 20% de données manquantes . On remarque que pour des pourcentage de donnés manquantes variant de 30% jusqu'a 60%, une légere augmentation des taux de RMSE entre 0.18 à 0.21 sur les deux datasets mais toujours de meilleur résulats sur la base de données massive "Diabètes". A partir de 60%, les taux de RMSE augmente de 0.21 jusqu'a arriver a 0.24 pour le Dataset "Breast" et 0.23 pour "Diabètes" qui contienne respectivement 80% de données manquantes. Des taux de RMSE très satisfaisant sur l'ensemble de la plage de valeurs manquantes avec des résultats plus précis sur la base de données massive Diabètes, puisque même si le pourcentage de données manquantes accroît il yaura toujours suffisamment de données pour un bon apprentissage par rapport a une petite base de données qui sera plus impactées par le pourcentage de données manquantes .

3.6.3 Comparaison FGAN avec d'autres méthodes apprentissage profond

Pour bien tester les performances et analyser les résultats du modèle proposer FGAN, il est nécessaire de le comparer à d'autre modèles existants, nous avons choisi le GAIN et Auto-encodeur [55] comme repères puisqu'ils ont présenté aussi d'excellents résultats en question de précision et fiabilité de données sur différents datasets notamment sur le dataset (BREAST) utilisé. La figure ci-dessous représente les trois courbes du FGAN et GAIN et l'Auto-encodeur :

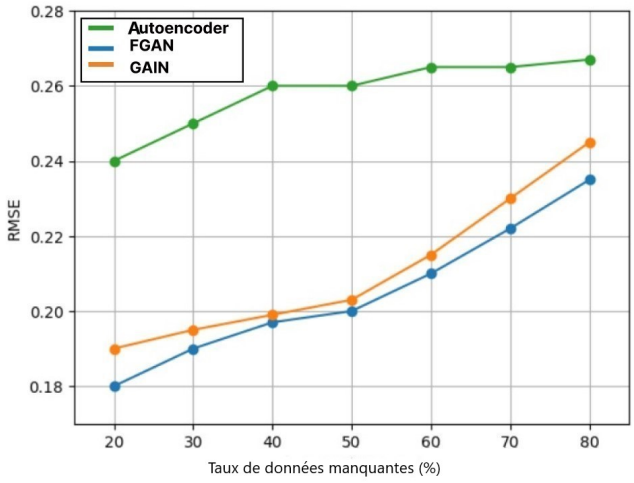

FIGURE 3.16 – Comparaison FGAN et GAIN et l'Auto-encodeur

La figue 3.20 montre que même si la performance des trois algorithmes diminue à mesure

que les taux des données manquantes augmentent, FGAN et GAIN surpasse systématiquement les indices de référence de l'Auto-encodeur sur l'ensemble de la gamme des taux manquants de 20% a 80% avec un très grand écart. On remarque aussi que les deux réseaux antagonistes génératifs[54] FGAN et GAIN ont des résulats stables de RMSE sur l'intervalle [0.18,0.20] pour des taux manquants de 20% jusqu'a 50% avec de meilleur résultat de précisions pour notre modèle FGAN proposé, a partir de 50% une légère augmentation de l'erreur RMSE de 0.25 pour GAIN et 0.23 pour le FGAN ce qui représente un dépassement performance très importante de 5% des taux de précisons par rapport au modèle GAIN. A partir de ces résultats, on peut dire que Les performances obtenues avec le modèle proposé auront un impact conséquent dans le domaine médical secteur qui exige un taux d'erreur minimal.

3.6.4 Comparaison FGAN avec d'autres méthodes apprentissage automatique

On a utilisé comme repère la méthode MissForest [72] la plus compétitive en tant que méthode d'imputation de données manquantes : cette méthode fonctionne avec l'algorithme Random Forest [40]. Il a surpassé tous les autres algorithmes dans tous les indicateurs, y compris KNN-Impute [39], dans certains cas de plus de 50% de données manquantes [72]. Dans l'article [55] MissForest a été aussi appliqué sur le Dataset (BREAST) utilisé. La figure ci-dessous représente la comparaison de MissForest avec notre modèle proposé FGAN :

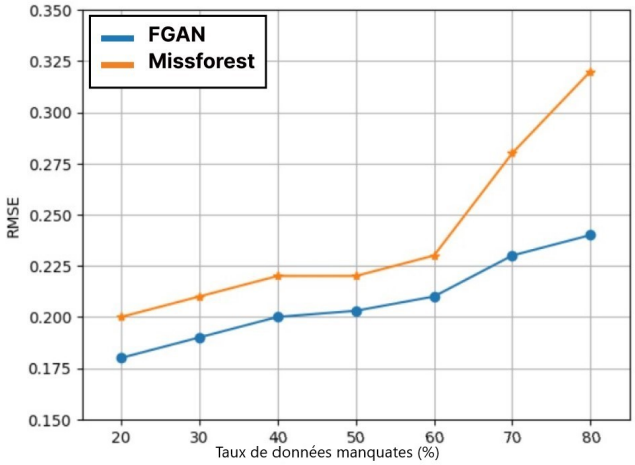

FIGURE 3.17 – Comparaison FGAN et MissForest

On remarque que notre modèle proposé FGAN donne d'excellent résultats avec des taux variant de données manquantes, comparativement a la méthodes MissForest.

3.7 Conclusion :

Dans ce chapitre, nous avons exposé notre modèle proposé pour l'imputation des données manquantes. Les étapes du modèle ont été ensuite détaillé. L'implémentation a permis de présenter les resulats obtenus qui s'averent trés satisfaisants. Une comparaison a été effectuée et a permis de montrer la grande performance du modèle proposé. Ces résultats seront souligner dans la conclusion générale avec quelques suggestions pour des travaux futurs.

Conclusion

Les méthodes d' apprentissage profond (Deep learning) sont de plus en plus utilisées pour la résolution des problèmes de "Missing Data". Elles s'appuient principalement sur l'analyse statistique et l'exploration de données. Ces méthodes sont basées sur des techniques exploratoires et des algorithmes pour découvrir les relations qui relient les données afin de fournir des résultats fiables. L'application de ces méthodes nous permet de mieux comprendre les données qui nous entourent et d'améliorer les performances pour prédire des résultats.

Dans ce travail, un modèle Fuzzy K-Means (FKM) amélioré par la méthode Elbow, a été proposé avec une méthode « Feature Selection ». Cette hybridation a permis d'apporter une amélioration aux résultats de classification. Ensuite, le réseau antagoniste génératif a été proposé pour prédire les valeurs des données manquantes. Ces résultats sont très prometteur surtout dans le domaine médical où l'information doit être complète pour la prise de décision.

Dans ce projet :
1. Un état de l'art sur les concepts du « Big Data » et des Datas médical a été présenté dans le premier chapitre.
2. Une étude des méthodes utilisées pour le traitement des données manquantes dans les « Big Data » médical a été exposée dans le deuxième chapitre.
3. Dans le chapitre 3
— L'implémentation de la méthode de partitionnement Fuzzy K-Means (FKM) amélioré par la méthode Elbow et « Feature Selection » ont été proposées.
— L'algorithme a été combiné avec le réseau antagoniste génératif et validé sur des données massives réelles « Médicales ».

— L'utilisation du Framework Spark a été d'un grand apport dans le traitement de ces données massives.
— La discussion et l'analyse des résultats obtenus par notre modèle d'imputation proposée (FGAN) à permis de souligner l'ampleur des performances remarquable du modèle et sa robustesse vis à vis à des taux de données manquantes croissantes.

En perspective, ce travail peut être étendu en prenant en considération les points suivants :
(1) Automatisation du choix des hyperparamètres des méthodes utilisées.
(2) Utilisation d'autre Datasets pour la valider du système proposé.
(3) Implémentation d'autres types de réseau tels le réseau"Convolutional Neural Network (CNN)" pour générer les données manquantes et developper une étude comparative avec le modèle proposé.
(4) Proposer un système pour la génération de données images (telles les images radiologiques) et qui présente une mauvaise qualité de résolution.

Bibliographie

[1] Soumojit Bose. What ways can Big Data Analytics make advertising more impactful ? . https://www.linkedin.com. Accessed : 10-12-22.

[2] « Marjolaine Tasset JDN , Le volume de données mondial sera multiplié par 45 entre 2020 et 2035 ». https://www.journaldunet.com. Accessed : 20-11-22.

[3] Hadoop MapReduce Tutorial. https://www.projectpro.io/hadoop-tutorial/hadoop-mapreduce-tutorial-. Accessed : 10-11-22.

[4] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1) :1–37, 2021.

[5] Comprendre les RDD pour mieux Développer en Spark. https://www.data-transitionnumerique.com/comprendre-rdd-spark/. Accessed : 17-3-23.

[6] Hayet Medfouni and Bilel Khantoul. Validation de clustering des donneés dans un contexte big data. 2018.

[7] Olivier JOUANNOT. Présentation générale big data. *Guide Share France*, 2013.

[8] Zachary A Vesoulis, Ameena N Husain, and F Sessions Cole. Improving child health through big data and data science. *Pediatric research*, 93(2) :342–349, 2023.

[9] Monerah Al-Mekhlal and Amir Ali Khwaja. A synthesis of big data definition and characteristics. In *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pages 314–322. IEEE, 2019.

[10] Babak Yadranjiaghdam, Nathan Pool, and Nasseh Tabrizi. A survey on real-time big data analytics : applications and tools. In *2016 international conference on computational science and computational intelligence (CSCI)*, pages 404–409. IEEE, 2016.

[11] Hausmane Issarane. Les 6 V du Big data . https://brightcape.co/6v-bigdata/, 02 2019. Accessed : 20-11-22.

[12] Manuel B Garcia. Cooperative learning in computer programming : A quasi-experimental evaluation of jigsaw teaching strategy with novice programmers. *Education and Information Technologies*, 26(4) :4839–4856, 2021.

[13] Arsia Amir-Aslani and Ricky Bhajun. Les 7 «v» piliers du big data. 11 2016.

[14] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1) :37–43, 2019.

[15] Manuel Au-Yong-Oliveira, Antonio Pesqueira, Maria José Sousa, Francesca Dal Mas, and Mohammad Soliman. The potential of big data research in healthcare for medical doctors' learning. *Journal of Medical Systems*, 45 :1–14, 2021.

[16] Liang Hong, Mengqi Luo, Ruixue Wang, Peixin Lu, Wei Lu, and Long Lu. Big data in health care : Applications and challenges. *Data and Information Management*, 2(3) :175–197, 2018.

[17] Jérôme Béranger. La valeur éthique des big data en santé. *Les Cahiers du numérique*, 12(1) :109–132, 2016.

[18] Jennifer Bresnick. Top 10 challenges of big data analytics in healthcare. *Health IT Analytics*, 2017.

[19] Liang Hong, Mengqi Luo, Ruixue Wang, Peixin Lu, Wei Lu, and Long Lu. Big data in health care : Applications and challenges. *Data and information management*, 2(3) :175–197, 2018.

[20] Relwende Aristide Yameogo. *Risques et perspectives du big data et de l'intelligence artificielle : approche éthique et épistémologique*. Theses, Normandie Université, September 2020.

[21] Apache Hadoop. https://hadoop.apache.org. Accessed : 3-11-22.

[22] Lala Septem Riza, Muhammad Naufal Fazanadi, Judhistira Aria Utama, Khyrina Airin Fariza Abu Samah, Taufiq Hidayat, and Shah Nazir. Sax and random projection algorithms for the motif discovery of orbital asteroid resonance using big data platforms. *Sensors*, 22(14) :5071, 2022.

[23] Apache Spark. https://spark.apache.org. Accessed : 3-11-22.

[24] Ravi Kanth Motupalli. Challenges, research issues and tools involved in big data and hadoop.

[25] Apache Storm. `https://storm.apache.org`. Accessed : 3-11-22.

[26] Ana I Torre-Bastida, Josu Díaz-de Arcaya, Eneko Osaba, Khan Muhammad, David Camacho, and Javier Del Ser. Bio-inspired computation for big data fusion, storage, processing, learning and visualization : state of the art and future directions. *Neural Computing and Applications*, pages 1–31, 2021.

[27] Bhupesh Rawat, Nidhi Mehra, Ankur Singh Bist, Muhamad Yusup, and Yulia Putri Ayu Sanjaya. Quantum computing and ai : Impacts & possibilities. *ADI Journal on Recent Innovation*, 3(2) :202–207, 2022.

[28] Sami Mahfoudhi, Ines Rahmany, Mushira Freihat, and Tarek Moulahi. Missing data recovery in the e-health context based on machine learning models. 2022.

[29] Eric V Strobl, Shyam Visweswaran, and Peter L Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6 :47–62, 2018.

[30] Pooja Rani, Rajneesh Kumar, and Anurag Jain. Hioc : a hybrid imputation method to predict missing values in medical datasets. *International Journal of Intelligent Computing and Cybernetics*, 14(4) :598–616, 2021.

[31] Luke Oluwaseye Joel, Wesley Doorsamy, and Babu Sena Paul. A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 5(3) :971–1005, 2022.

[32] Zhongheng Zhang. Multiple imputation for time series data with amelia package. *Annals of translational medicine*, 4(3), 2016.

[33] Jeeyae Choi, Jeungok Choi, and Hee-Tae Jung. Applying machine-learning techniques to build self-reported depression prediction models. *CIN : Computers, Informatics, Nursing*, 36(7) :317–321, 2018.

[34] Andrew J Steele, Spiros C Denaxas, Anoop D Shah, Harry Hemingway, and Nicholas M Luscombe. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one*, 13(8) :e0202344, 2018.

[35] Kevin Kunzmann, Lorenz Wernisch, Sylvia Richardson, Ewout W Steyerberg, Hester Lingsma, Ari Ercole, Andrew IR Maas, David Menon, and Lindsay Wilson. Imputation of ordinal outcomes : a comparison of approaches in traumatic brain injury. *Journal of neurotrauma*, 38(4) :455–463, 2021.

[36] Approach to handle missing values in different dataset. 22(6) :0048–4911, 2022.

[37] Sebastian Daberdaku, Erica Tavazzi, and Barbara Di Camillo. A combined interpolation and weighted k-nearest neighbours approach for the imputation of longitudinal icu laboratory data. *Journal of Healthcare Informatics Research*, 4(2) :174–188, 2020.

[38] Mark J Giganti, Pamela A Shaw, Guanhua Chen, Sally S Bebawy, Megan M Turner, Timothy R Sterling, and Bryan E Shepherd. Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples, and multiple imputation. *The annals of applied statistics*, 14(2) :1045, 2020.

[39] Sebastian Daberdaku, Erica Tavazzi, and Barbara Di Camillo. A combined interpolation and weighted k-nearest neighbours approach for the imputation of longitudinal icu laboratory data. *Journal of Healthcare Informatics Research*, 4 :174–188, 2020.

[40] Bokai Yang, Guangzhe Dai, Yujie Yang, Darong Tang, Qi Li, Denan Lin, Jing Zheng, and Yunpeng Cai. Automatic text classification for label imputation of medical diagnosis notes based on random forest. In *Health Information Science : 7th International Conference, HIS 2018, Cairns, QLD, Australia, October 5–7, 2018, Proceedings 7*, pages 87–97. Springer, 2018.

[41] Solomiia Fedushko, Taras Ustyianovych, et al. Medical card data imputation and patient psychological and behavioral profile construction. *Procedia Computer Science*, 160 :354–361, 2019.

[42] Karima BENHAMZA, Nadjette BENHAMIDA, Mohamed Ilyes BOURAHDOUN, and Bilel BOUDJAHEM. Hybrid analytic method for missing data imputation in medical big data. *International Journal of Informatics and Applied Mathematics*, 5(2) :1–11.

[43] Hufsa Khan, Xizhao Wang, and Han Liu. Handling missing data through deep convolutional neural network. *Information Sciences*, 595 :278–293, 2022.

[44] Himansu Das, Bighnaraj Naik, HS Behera, Shalini Jaiswal, Priyanka Mahato, and Minakhi Rout. Biomedical data analysis using neuro-fuzzy model with post-feature

reduction. *Journal of King Saud University-Computer and Information Sciences*, 34(6) :2540–2550, 2022.

[45] Huimin Wang, Jianxiang Tang, Mengyao Wu, Xiaoyu Wang, and Tao Zhang. Application of machine learning missing data imputation techniques in clinical decision making : taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*, 22(1) :1–14, 2022.

[46] Hiroaki Tomita, Hironori Fujisawa, and Masayuki Henmi. A bias-corrected estimator in multiple imputation for missing data. *Statistics in Medicine*, 37(23) :3373–3386, 2018.

[47] Shu-Fen Huang and Ching-Hsue Cheng. A safe-region imputation method for handling medical data with missing values. *Symmetry*, 12(11) :1792, 2020.

[48] Tomasz Szandała. Review and comparison of commonly used activation functions for deep neural networks. *Bio-inspired neurocomputing*, pages 203–224, 2021.

[49] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[50] Jason Brownlee. A gentle introduction to cross-entropy for machine learning. *Machine Learning Mastery. https ://machinelearningmastery. com/cross-entropy-for-machine-learning*, 2019.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[52] Jale Bektaş, Turgay Ibrikçi, and İsmail Türkay Özcan. The impact of imputation procedures with machine learning methods on the performance of classifiers : An application to coronary artery disease data including missing values. *Biomedical Research*, 29(13) :2780–2785, 2018.

[53] Chung-Yuan Cheng, Wan-Ling Tseng, Ching-Fen Chang, Chuan-Hsiung Chang, and Susan Shur-Fen Gau. A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder. *Frontiers in psychiatry*, 11 :673, 2020.

[54] Jaeyoon Kim, Donghyun Tae, and Junhee Seok. A survey of missing data imputation using generative adversarial networks. In *2020 International conference on artificial intelligence in information and communication (ICAIIC)*, pages 454–456. IEEE, 2020.

[55] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain : Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

[56] Hongyang Zhang and David P Woodruff. Medical missing data imputation by stackelberg gan. *Carnegie Mellon University*, 2018.

[57] Diogo Telmo Neves, Marcel Ganesh Naik, and Alberto Proença. Sgain, wsgain-cp and wsgain-gp : Novel gan methods for missing data imputation. In *Computational Science–ICCS 2021 : 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part I*, pages 98–113. Springer, 2021.

[58] Yuan Luo. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1) :bbab489, 2022.

[59] Amelia Ritahani Ismail, Nadzurah Zainal Abidin, and Mhd Khaled Maen. Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control (JRC)*, 3(2) :143–152, 2022.

[60] Peter C Austin, Ian R White, Douglas S Lee, and Stef van Buuren. Missing data in clinical research : a tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37(9) :1322–1331, 2021.

[61] Hanane NAIDJA Karima BENHAMZA. GANs Models for Medical Missing Data Imputation, Conference IAM university 8 May 1945 Guelma, 2022.

[62] Chia-Hui Liu, Chih-Fong Tsai, Kuen-Liang Sue, and Min-Wei Huang. The feature selection effect on missing value imputation of medical datasets. *Applied Sciences*, 10(7) :2344, 2020.

[63] Variance inflation factor. `www.statsmodels.orginfluence.variance_inflation_factor.html`. Accessed : 20-06-23.

[64] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics : Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

[65] Akanksha Kapoor and Abhishek Singhal. A comparative study of k-means, k-means++ and fuzzy c-means clustering algorithms. In *2017 3rd international conference on computational intelligence & communication technology (CICT)*, pages 1–6. IEEE, 2017.

[66] MA Syakur, BK Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series : materials science and engineering*, volume 336, page 012017. IOP Publishing, 2018.

[67] Oskar Morgenstern John von Neumann. *Theory of Games and Economic Behavior*. 1944.

[68] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[69] Tim van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7) :3797–3820, 2014.

[70] Shraddha Pandit, Suchita Gupta, et al. A comparative study on distance measuring approaches for clustering. *International journal of research in computer science*, 2(1) :29–31, 2011.

[71] Karen Grace-Martin. Assessing the fit of regression models. *The Analysis Factor*, 2015, 2018.

[72] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1) :112–118, 2012.