N^o d'ordre : ..../..../....

الجمهـــورية الجـــزَائرية الديمـــقراطية الشّعـــبية

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزَارة التـعليم العـالي وَ البـحث العـلمي

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

جـامعة محـمد خيضـر بـسكـرة

UNIVERSITÉ MOHAMED KHIDER BISKRA

كليـة العلـوم الدّقيقـة وَ علـوم الطّبيعـة وَ الحيـاة

FACULTÉ DES SCIENCES EXACTES ET DES SCIENCES DE LA NATURE ET DE LA VIE

قسـم الإعـلام الآلـي

DÉPARTEMENT D'INFORMATIQUE

مخبـر الذّكـاء المعـلومـاتي

LABORATOIRE DE L'INFORMATIQUE INTELLIGENTE (LINFI)

THÈSE

Présentée en vue de l'obtention du diplôme de

Doctorat 3ème cycle (LMD) en Informatique

Option : **Intelligence Artificielle**

Par : **Djihane HOUFANI**

Titre

Une approche basée sur les SMA et méta-heuristique pour la prédiction (PHM) médicale

---

Soutenue publiquement le : ../../.... devant le jury composé de:

Président	*Pr. Khaled REZEG*	*Université de Biskra*
Rapporteur	*Pr. Sihem SLATNIA*	*Université de Biskra*
Co-Rapporteur	*Pr. Okba KAZAR*	*Université de Biskra*
Examinateur	*Pr. Saber BENHARZALLAH*	*Université de Batna 2*
Examinateur	*Dr. Djamel NESSAH*	*Université de Khenchela*
Examinatrice	*Pr. Rachida SAOULI*	*Université de Biskra*

Année universitaire: 2022-2023

Remerciements

Tout d'abord, je tiens à remercier Dieu, le tout puissant et miséricordieux, qui m'a donnée la force et la patience pour accomplir ce travail.

J'adresse ma profonde gratitude à ma Directrice de thèse Madame ***Sihem SLATNIA****, Professeur à l'Université de Biskra. Un grand merci pour sa disponibilité, son soutien et ses conseils.*

J'adresse ma profonde reconnaissance à Monsieur ***Okba KAZAR****, Professeur à l'Université de Biskra, qui m'a dirigée tout au long de la préparation de ce travail. Un grand merci pour sa disponibilité, sa confiance et ses conseils, nombreux et avisés.*

Je tiens à remercier Madame ***Guadalupe ORTIZ BELLOT****, Professeur au Laboratoire UCASE, Cádiz en Espagne pour sa louable collaboration.*

Mes sincères remerciements vont au Professeur ***Khaled REZEG*** *de* ***l'Université de Biskra****, qui a accepté de présider mon Jury de thèse, ainsi que le Professeur* ***Rachida SAOULI*** *de* ***l'Université de Biskra****, le Professeur* ***Saber BENHARZALLAH*** *de* ***l'Université de Batna*** *et le Docteur* ***Djamel NESSAH*** *de* ***l'Université de Khenchela*** *qui ont gentiment accepté de participer à mon Jury.*

J'adresse mes remerciements aux Docteurs ***Hamza SAOULI****,* ***Meftah ZOUAI****, et* ***Abdelhak MERIZIG*** *et mes amis* ***Ikram REMADNA*** *et* ***Salah Eddine HENOUDA*** *pour leur aide et leurs conseils pertinents.*

Enfin, Je tiens à remercier tout le cadre professionnel et administratif de ***la Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie****, et du* ***Département de l'Informatique*** *en particulier.*

Djihane

Dédicaces

A mes chers parents, pour tous les sacrifices consentis, leur tendresse, leur soutien et leurs prières pour atteindre cette étape de ma vie.

A mon cher frère, Djaouad pour ses encouragements permanents et son soutien moral.

A une personne très spéciale, Imad pour son soutien, son encouragement, sa précieuse aide et sa confiance.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire, que ce travail soit l'accomplissement de leurs vœux tant allégués.

A tous mes amis et tous mes collègues du Laboratoire d'INFormatique Intelligente (LINFI) de l'Université de Biskra.

A la mémoire de mon oncle Abdelatif et mon grand-père Mohamed Saleh, qui nous ont quitté l'année passée.

Je dédie ce travail.

Résumé

Au cours des dernières années, l'intelligence artificielle (IA) ne cesse de révolutionner le monde. Elle est intégrée dans plusieurs secteurs tels que l'économie, l'industrie, la biologie, la médecine, etc. L'utilisation de l'IA dans en médecine présente un grand intérêt pour les chercheurs qui exploitent l'approche prédictive pour son importance dans la prise de décision. Elle ouvre des perspectives prometteuses pour améliorer la qualité des soins au profit du patient à travers une prise en charge personnalisée, une bonne détection des symptômes et une exploitation des résultats d'analyse (imagerie médicale, rapports médicaux, tests sanguins,etc.) pour un meilleur diagnostic. Cependant, malgré l'impact positif de l'IA sur le secteur sanitaire, elle est confrontée à de nombreux challenges notamment, la manipulation des big data, la diversité des offres de soins, la durée du diagnostic, la complexité et la hausse des coûts de traitement.

La médecine prédictive vise à déterminer la probabilité d'atteindre une maladie, à prédire la récidive, le taux de létalité et la propagation d'une maladie dans une zone. L'application des technologies telles que la biotechnologie, la génomique et les techniques de l'IA (IoT, SMA, apprentissage automatique, etc.) peuvent améliorer la modélisation distribuée des systèmes intelligents et les méthodes de classification. L'objectif principal de cette étude est de proposer une approche intelligente pour :

- améliorer la qualité du diagnostic médical et la détection des pathologies en permettant d'éviter de soumettre les patients à des examens intrusifs ;
- définir des stratégies thérapeutiques plus adaptées aux situations des patients ;
- optimiser les parcours de soins (détection précoce, gain de temps, coûts, etc.).

Le cancer du sein est l'une des causes les plus fréquentes de décès chez la femme. De plus, au courant des trois dernières années, l'apparition de la pandémie Covid-19 a laissé un impact négatif sur des milliers de personnes dans le monde. Cette crise sanitaire a également donné un dur coup à l'évolution de l'économie au niveau mondial. L'application de nos objectifs pour la prise en charge de ces deux pathologies constitue un segment important de ce projet, sachant que la quantité et la qualité des données disponibles sont des éléments clés sur les applications de l'IA en santé.

Mots clés : *Méta-heuristique, système multi-agent, PHM, informatique prédictive, apprentissage automatique, cancer du sein, COVID-19.*

Abstract

In recent years, Artificial Intelligence (AI) has made a great revolutionary progress in the world. AI has been incorporated into several fields including the economic, industrial, biological, and medical sectors, etc. The integration of AI in the medical field presents a great interest to researchers who exploit the predictive approach due to its importance in decision-making. It offers promising prospects for improving the quality of care for the benefit of the patient through personalized and predictive care, good detection of symptoms for better diagnosis, and the use of analysis results (medical imaging, medical reports, tests blood, etc.). However, despite the positive impact of AI on the healthcare sector, it faces many challenges. In particular, the manipulation of big data, the growing complexity, the diversity of healthcare offers, the length of diagnosis, and the increase in treatment costs.

Predictive medicine aims to determine the probability of reaching a disease, to predict the recurrence, the case fatality rate, and disease forecasting. The application of biotechnology, genomics, and AI techniques (IoT, SMA, machine learning, and optimization techniques) can improve distributed intelligent modeling systems and classification methods. The main objective of this thesis is to propose a new intelligent approach for :

- improving the quality of medical diagnosis and the detection of pathologies by avoiding subjecting patients to intrusive examinations ;
- defining therapeutic strategies more suited to patient situations ;
- optimizing the care pathways (early detection, saving time, costs, etc.).

Breast cancer is one of the most common causes of death in women. In addition, over the past three years, the emergence of the new Covid-19 pandemic had a harmful impact worldwide. This health crisis has also taken a heavy toll on the development of the economy around the world. The application of our objectives for the management of these two pathologies constitutes an important segment of this research project, knowing that the quantity of available data and the quality of their annotation are key elements on the applications of AI in health.

Keywords : *Metaheuristics, multi-agent system, PHM, predictive informatics, machine learning, breast cancer, COVID-19.*

الملخص

حقق الذكاء الاصطناعي (AI) في السنوات الأخيرة تقدمًا ثوريًا كبيرًا في العالم. فقد تم دمجه في العديد من المجالات بما في ذلك القطاعات الاقتصادية والصناعية والبيولوجية والطبية ، وما إلى ذلك. يحظى استخدام الذكاء الاصطناعي في مجال الطب باهتمام كبير من طرف الباحثين الذين يستغلون النهج التنبئي لأهميته في صنع القرار. إنه يفتح آفاقًا واعدة لتحسين جودة الرعاية لصالح المريض من خلال الرعاية الشخصية والتنبؤية، والكشف الجيد عن الأعراض من أجل تشخيص أفضل واستخدام نتائج التحليل (التصوير الطبي، والتقارير الطبية ، والاختبارات الدم ... إلخ.). على الرغم من التأثير الإيجابي للذكاء الاصطناعي على قطاع الصحة ، فإن هذا الأخير يواجه العديد من التحديات. على وجه الخصوص، العمل على الكميات الهائلة من البيانات المتعددة النماذج و المصادر مما يجعل معالجتها أكثر تعقيدا، إضافة الى تنوع عروض الرعاية الصحية، ومدة التشخيص وزيادة تكاليف العلاج.

يهدف الطب التنبئي إلى تحديد احتمالية الوصول إلى مرض ما، والتنبؤ بتكرار حدوثه، ومعدل إماتة الحالة، وانتشار المرض في منطقة ما. تطبيق تقنيات مثل التكنولوجيا الحيوية وعلوم الجينوم وتقنيات الذكاء الاصطناعي (إنترنت الأشياء ، الأنظمة متعددة الوكلاء SMA ، تقنيات التعلم الآلي والتحسين) مكن من تحسين نمذجة الأنظمة الذكية الموزعة وتصنيف الأساليب.

الهدف الرئيسي من هذه الأطروحة هو اقتراح نظام ذكي يساهم في:

- تحديد استراتيجيات علاجية أكثر ملاءمة لحالات المريض .
- تحسين جودة التشخيص الطبي واكتشاف الأمراض من خلال تجنب إخضاع المرضى لفحوصات تدخلية.
- تعزيز مسارات الرعاية (الاكتشاف المبكر ، توفير الوقت ، التكاليف ، إلخ.).

يعد سرطان الثدي أحد أكثر أسباب الوفاة شيوعًا عند النساء. بالإضافة إلى ذلك ، على مدى السنوات الثلاث الماضية، ترك ظهور جائحة كوفيد -19 تأثيرًا سلبيًا على آلاف الأشخاص حول العالم. لقد ألحقت هذه الأزمة الصحية خسائر فادحة بتنمية الاقتصاد في جميع أنحاء العالم. يشكل تطبيق الأهداف المذكورة سابقا لمواجهة هذين المرضين جزءًا مهمًا من هذه الأطروحة، مع العلم أن كمية البيانات المتاحة جودتها و توفرها هي عناصر أساسية في تطبيقات الذكاء الاصطناعي في الصحة.

الكلمات المفتاحية:

الاستدلال الفوقي ، الأنظمة متعددة الوكلاء، PHM، الحوسبة التنبؤية، التعلم الآلي، سرطان الثدي ، كوفيد-19.

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

Liste des abréviations

AG : Algorithme Génétique

ADN : Acide Désoxyribonucléique

ANN : Artificial Neural Network

AR : Associative Rules

ARIMA : Auto Regressive Integrated Moving Average

ARN : Acide Ribonucléique

AUC : Area Under Curve

BC : Breast Cancer

BCC : Breast Cancer Coimbra dataset

BCDR : Breast Cancer Digital Repository

Bi-RADS : Breast Imaging-Reporting And Data System

BN : Bayesian Network

BPNN : Back Propagation Neural Network

BRF : Balanced Random Forest

CAD : Computer-Aided Diagnosis

CART : Classification And Regression Trees

CBR : Case-Based Reasoning

CCHS : Canadian Community Health Survey

CHDD : Cleveland Heart Disease Data set

CL : Competitive Learning

CNN : Convolutional Neural Network

CSSVM : Cost Sensitive Support Vector Machine

CT : Computed Tomography

CVD : CardioVascular Disease

CXR : Chest X-Ray Images

DBN : Deep Blief Network

DCE-IRM : Dynamic Contrast Enhanced Magnetic Resonance Imaging

DDSM : Digital Database for Screening Mammography

DE algorithm : Differential Evolution algorithm

DL : Deep Learning

EM : Expectation Maximization

FCN : Fully Convolutional Network

FL : Fuzzy Logic

FNAC : Fine-Needle Aspiration Cytology

FPR : False Positive Rate

FRS : Framingham Risk Score

FNAC : Fine-Needle Aspiration Cytology

FRS : Framingham Risk Score

GBM : Gradiant Boosting Machine

GD : Gradiant Descent

GDCNN : Genetic Deep Convolutional Neural Network

GONN : Genetically Optimized Neural Network

GPR : Gaussian Process Regression

GRU : Gated Recurrent Unit

HAS : Harmony Search

HMM : Hidden Markov Model

IA/AI : Intelligence Artificielle/ Artificial Intelligence

IDSS : Intelligent Decision Support System

IGSAGW : Information Gain Directed Simulated Annealing Genetic Algorithm Wrapper

IHC : Incomplete Heterogeneous COVID-19

IoT : Internet of Things

IR-FRE : Improved Random Forest -based Rule Extraction

IRRCNN : Inception Recurrent Residual Neural Network

KMC : k-Means Clustering

KNN : K-Nearest Neighbour

KNN V : K-Nearest Neighbour Variant

K-SVM : Kernel Support Vector Machine

LR : Logistic Regression

L-SVM : Linear Support Vector Machine

LSSOED : Life-Sensitive Self-Organizing Error Drive

LSTM : Long Short-Term Memory

LVQ : Learning Vector Quantization

MAD : Median Absolute Derivation

MAE : Mean Absolute Error

MCC : Matthews Correlation Coefficient

ME-CNN : Mixed Ensemble of Convolutional Neural Networks

MLP : Multi Layer Perceptron

MSE : Mean Squared Error

NB : Naive Bayes

NSGA-II : Non-dominated Sorting Genetic Algorithm II

OMS : Organisation Mondiale de la Santé

OSAS : Obscurtive Sleep Apnea Syndrome

PA : Prophet Algorithm

PCA : Principal Componant Analysis

PCR : Polymerase Chain Reaction

PHM : Prognostics and Health Management

PPV : Positive Predictive Value

PR : Precision-Recall

PSO : Practical Swarm Optimization

PSOWNN : Particle Swarm Optimized Wavelet Neural Network

QNN : Quantum Neural Network

RBF : Radial Basic Function

RBFN : Radial Basis Function Neural Network

RF : Random Forest

RMSE : Root-Mean-Square Error

ROC : Receiver Operating Characteristic

SEER : Surveillance, Epidemiology, and End Results database

SEM : Structural Equation Model

SGD : Stochastic Gradient Descent

SMA/MAS : Système multi agents/ Multi agent system

SOM : Self Organizing Maps

SSL : Spine Surgey Likelihood

SVCMAC : Self-Validation Cerebellar Model Articulation Controller

SVM : Support Vector Machine

TPR : True Positive Rate

VGG 16 : Visual Geometry Group 16

WBC : Wisconsin Breast Cancer

WBCP : Wisconsin Breast Cancer Prognosis

WDBC : Wisconsin Diagnosis Breast Cancer

WOBC : Wisconsin Original Breast Cancer

Introduction Générale

Contexte du travail

Depuis l'apparition de l'Intelligence Artificielle (IA) dans les années cinquante [1], son but ultime était de simuler le comportement humain dans l'apprentissage et la prédiction de situations futures. Les scientifiques et les chercheurs du monde entier, étaient enthousiasmés par le progrès des innovations, résultant d'un besoin naturel de créer des technologies plus avancées. Ces innovations peuvent permettre à l'humanité d'atteindre des avancées majeures dans tous les domaines. L'IA permet en grande partie de stocker et de traiter de grandes quantités de données de manière intelligente et spécifiquement de transformer ces données en informations, qui pourraient être utilisées comme outils pratiques. Elle a attiré de nombreux utilisateurs depuis sa naissance, on la trouve dans les entreprises, les recherches scientifiques et le domaine médical, etc. Aujourd'hui, sa transition progressive dans le domaine médical permet d'exploiter des systèmes basés sur ses différentes techniques pour assurer des pratiques médicales délicates telles que le diagnostic, le pronostic, la proposition de traitement selon l'état de santé du patient, etc. Ces systèmes donnent des résultats plus précis et évitent les erreurs humaines.

La médecine prédictive consiste à déterminer la probabilité d'une maladie, dont le rôle principal est de diminuer son impact sur le patient, par exemple en prévenant la mortalité ou en limitant la morbidité. Malgré les nombreuses solutions proposées dans la littérature, la prédiction médicale reste une tâche difficile qui demande beaucoup d'efforts. Ceci est attribué à son importance vitale dans la prise de décision. Les principaux objectifs de la médecine prédictive sont : (a) La collecte des données médicales [2] ; (b) l'analyse de ces données pour prédire les risques que confrontent le patient ; (c) prédire l'efficacité d'un traitement donné sur les individus, puis intervenir avant la production du résultat.

L'informatique médicale est à la jonction de la médecine, des outils informatiques et de l'IA. Cette discipline fournit de multiples services ; elle est utilisée pour optimiser l'exactitude, l'efficacité des systèmes et améliorer la santé publique. Elle est appliquée également dans la maintenance de la confidentialité et la sécurité des données médicales des patients. Le but ultime de cette discipline est de développer des systèmes pour assister les médecins pour une meilleures prise en charge de leurs patients. De tels systèmes devraient réduire les erreurs médicales, accélérer les interventions cliniques et fournir de meilleurs soins, en disposant des informations requises au bon endroit et au bon moment. Le développement de ces systèmes repose sur deux principaux critères :

— **Les données médicales :**

La collecte et l'analyse des données médicale font partie des tâches les plus coûteuses en termes de temps et de complexité. Elles sont générées massivement sous différents types (chiffres, images, textes, vidéos) et de diverses sources médicales, telles que les informations sur les patients, les biomarqueurs (génomique, protéomique, etc.), les résultats des diagnostics (tests sanguins, radiologie, etc.), les données statistiques (données sur les coûts et les réclamations, données sur la population et la santé publique) et sur le comportement (données provenant d'applications mobiles, de médias sociaux, de capteurs, d'appareils portables et de moniteurs de tension[3]). Le volume de ces données, leur hétérogénéité, leur privatisation et leur complexité, constituent un défi majeur pour leur traitement.

— **Les outils et méthodes de développement :**

Le choix de la bonne technique dépend souvent du problème à résoudre, de l'objectif de l'étude, de la représentation et du stockage des données. Il a un impact vital sur la mise en œuvre de tels systèmes, leur robustesse et leur précision.

L'apprentissage automatique, l'apprentissage en profondeur, les SMA et les approches d'optimisation montrent des résultats promoteurs dans des tâches dites difficiles auparavant, notamment l'analyse d'images, le traitement du langage, la recherche d'informations et la prévision. Ces techniques de l'IA, sont bien adaptées aux données médicales, car elles permettent d'identifier des modèles sur des données clairsemées, bruitées et hétérogènes. Les résultats actuels de l'application de ces techniques dans le développement des systèmes intelligents d'aide à la décision (IDSS) ont montré que, dans certains cas, les performances surpassent les médecins et les experts, pour diagnostiquer et prédire différentes maladies. Néanmoins, ces résultats doivent encore être améliorés,

pour établir un diagnostic précis. Dans cette thèse, nous allons traiter quelques problèmes rencontrés lors du développement des IDSS, plus particulièrement, ceux liés à la qualité de la décision.

Problématique et objectifs

La médecine est confrontée à de nombreux défis. En particulier, la gestion des Big Data, qui est un problème critique en raison de sensibilité de ces données. De plus, leur croissance exponentielle et leur accès limité les rendent parfois plus complexes et augmentent le temps et le coût du diagnostic. La confidentialité et la sécurité sont également des enjeux majeurs auxquels le domaine médical est confronté. Ces défis ont un impact négatif sur la plupart des fournisseurs sanitaires et des patients. L'IA, dans ce cas, crée un nouvel horizon d'analyse, qui combine des approches réalistes et informatiques, avec les connaissances des professionnels pour fournir de nouvelles innovations. Son but principal est de former des outils de plus en plus utiles et de saisir les principes qui rendent l'intelligence accessible. La médecine est un domaine qui peut bénéficier de cette opportunité, en améliorant la qualité des données et des pratiques médicales au profit des patients et des fournisseurs de soins.

L'intégration de l'IA dans la santé, est actuellement un sujet de grand intérêt pour l'importance vitale qu'elle apporte dans la prise de décision. Cependant, les données médicales sont de nature très hétérogène et nécessitent une phase de prétraitement pour entraîner des modèles prédictifs. L'intelligence artificielle en santé a des perspectives prometteuses pour améliorer la qualité des soins au profit des patients ; elle permet une bonne détection des symptômes pour un meilleur diagnostic. Elle permet également d'exploiter les résultats d'analyses (imagerie médicale...), de définir de nouvelles hypothèses diagnostiques et de formuler des propositions thérapeutiques plus personnalisées. Les systèmes multi-agents et les techniques de méta-heuristique pour l'optimisation peuvent améliorer respectivement les méthodes de modélisation et de classification des systèmes intelligents distribués. Plusieurs travaux ont été menés dans cette thématique et plusieurs systèmes d'aide à la décision ont été construits. Cependant, peu de ces travaux de recherche ont effectivement été intégrés à la pratique clinique. Cela est causé par la méfiance de la société et des médecins envers les structures, les logiciels et les produits de santé, en plus de la complexité du comportement imprévisible des systèmes biologiques et la sensibilité de travailler sur des vies humaines.

L'idée principale de ce travail de recherche est de proposer une nouvelle approche prédictive basée sur des méthodes de l'IA, notamment les méta-heuristiques, les systèmes multi-agents et les techniques de l'apprentissage artificiel pour la prédiction médicale et la prise de décision en temps réel. Le processus décisionnel de l'agent sera optimisé par les méta-heuristiques pour prédire une situation complexe ou un comportement évolutionnaire. Nos principaux objectifs sont :

- Progresser dans la détection des pathologies et éviter d'exposer les patients à des examens intrusifs.
- Travailler sur des données volumineuses de différentes sources.
- Élaborer un diagnostic et une stratégie thérapeutique adaptée aux besoins du patient, à son environnement et à son mode de vie.
- Prédire la réponse d'un patient à son traitement.

Contributions

En se basant sur les problèmes et les défis cités précédemment, une étude comparative et deux contributions sont proposées :

- **Une revue de la littérature de plusieurs travaux, en rapport avec notre thématique de recherche**

  Un bilan détaillé de plusieurs études menées durant la dernière décennie, en rapport avec notre thématique est fourni. En parcourant la revue de la littérature, nous avons remarqué que les chercheurs s'intéressent à la prédiction médicale, notamment le diagnostic et le pronostic de plusieurs pathologies (cancer du sein, maladies cardiovasculaires, Covid-19, etc.). Ils ont appliqué des méthodes et des approches de l'intelligence artificielle, telles que l'ANN, l'apprentissage en profondeur et le data mining, etc. Leurs résultats ont montré l'efficacité en termes d'exactitude. Cependant, la plupart des systèmes proposés prennent du temps dans la phase d'apprentissage. On peut aussi remarquer que très peu de ces travaux de recherche ont effectivement été intégrés à la pratique clinique.
- **Une étude comparative des méthodes de l'apprentissage artificiel pour la classification des tumeurs mammaires**

  Le cancer du sein est l'une des principales causes de décès chez les femmes dans le monde.

C'est le deuxième type de cancer qui touche les femmes après le cancer du poumon. Son diagnostic est un processus coûteux en termes de temps et peut comporter des erreurs. Les techniques d'apprentissage automatique sont largement utilisées pour améliorer l'efficacité du diagnostic médical et la prise de décision.

Pour choisir la méthode d'apprentissage artificielle que nous allons utiliser pour notre première contribution, nous avons effectué une étude comparative de la performance de différentes techniques de ML : K-SVM, L-SVM, LR, DT, des k- NN, RF et MLP pour le diagnostic du cancer du sein en utilisant la base de données publique WDBC. Les résultats expérimentaux prouvent l'efficacité des méthodes MLP et LR dans la classification du cancer du sein avec une exactitude de 98%.

- **Diagnostic assisté par ordinateur pour la classification des tumeurs mammaires**

  Pour cette première approche, nous avons proposé un CAD pour la détection du cancer du sein basé sur la méthode PCA et la régression logistique sur les ensembles de données WDBC et WOBC. Elle peut être résumée comme suit :

  — Réduction de la dimensionnalité des données en utilisant la méthode PCA.

  — Classification binaire des tumeurs BC, en appliquant la méthode d'apprentissage automatique par régression logistique.

  Les résultats obtenus révèlent que la combinaison de la méthode de réduction des caractéristiques PCA et la LR a augmenté la performance du diagnostic avec une exactitude de 100% et de 97% pour WDBC et WOBC respectivement. Cette combinaison a amélioré la qualité des données et a réduit le temps de calcul.

- **Un système prédictif intelligent basé SMA, DL, et méta-heuristiques pour la prédiction de la COVID-19**

  La pandémie de COVID-19 est une menace sanitaire mondiale et elle s'est propagé dans le monde entier. Elle a affecté tous les secteurs avec des décisions gouvernementales inédites de confinements et des protocoles sanitaires stricts. Pour faire face à l'augmentation exponentielle des contaminations et des décès à travers le monde, la communauté scientifique a fourni des efforts colossaux de recherches dans l'identification du virus mis en cause, la classification des symptômes, le diagnostic, le dépistage, et la recherche de vaccin en appliquant les technologies de l'IA.

Notre seconde contribution consiste à développer un système d'aide à la décision basé sur le paradigme SMA, une nouvelle technique de DL (TabNet) et méta-heuristiques (NSGA-II) pour la prédiction de l'admission des patients atteints de la COVID-19 aux soins intensifs, en utilisant une base de données publique disponible sur Kaggle.

Structure de la thèse

Cette thèse est organisée comme suit :

L'introduction générale définit le contexte, la problématique, les objectifs et les principales contributions de la thèse.

Le chapitre I présente brièvement le cadre théorique lié à notre thématique de recherche. La première partie introduit le concept d'agent et de SMA en présentant les différentes interactions et les types de communication dans un SMA et leur application dans la médecine. La deuxième partie, consiste à présenter l'apprentissage automatique de façon détaillée. La dernière partie présente un état de l'art sur les méta-heuristiques pour la résolution de problèmes d'optimisation.

Le chapitre II définit la médecine 4P et le PHM, présente une vue d'ensemble sur les contributions proposées dans la littérature pour la prédiction médicale. Ensuite, une synthèse bibliographique détaillée est élaborée. Enfin, les enjeux de l'application de l'IA pour la prédiction médicale sont présentés.

Le chapitre III aborde nos contributions qui consistent en :

- Une étude comparative expérimentale des différentes méthodes de l'apprentissage artificiel appliquées à la classification des tumeurs mammaires.
- Une approche basée sur la méthode d'extraction de caractéristiques (PCA) et la méthode d'apprentissage automatique (régression logistique), pour le diagnostic du cancer du sein.
- Une approche basée SMA, DL et méta-heuristique, pour la prédiction de la Covid-19.

En guise de **conclusion**, nous présenterons une synthèse des contributions, et nous énumérons nos perspectives et nos futures orientations en se basant sur les travaux réalisés.

Chapitre 1

Contexte de travail et concepts de base

1.1 Introduction

Ce chapitre présente brièvement les concepts de base liés à notre thématique de recherche. Dans la première partie, nous allons introduire le concept **"Agents"** à travers sa définition, ses caractéristiques et ses types. Ensuite, nous allons définir les **"Systèmes Multi-Agent"** et présenter les différentes interactions et les types de communication au sein d'un SMA. Enfin, nous allons aborder l'application des SMA dans le domaine médical. La seconde partie est une présentation détaillée de **l'apprentissage automatique**. Dans la troisième et dernière partie, un état de l'art sur **les métaheuristiques** et **l'optimisation** sera présenté.

1.2 Agents et Systèmes Multi-Agents

Les concepts agent et SMA sont relativement récents en Informatique. Leur apparition est initiée par les recherches en robotique et elle est issue de l'Intelligence Artificielle Distribuée (IAD) [1]. Les SMA regroupent des entités artificielles (des programmes, des robots) en interaction les unes avec les autres et avec l'environnement.

1.2.1 La notion d'agent

Le terme agent est utilisé dans diverses applications par des communautés venant de différents horizons. Nous présentons ici les concepts les plus adoptés.

1.2.1.1 Définition de l'agent

La définition d'un agent est bien évidemment un point essentiel dans la conception d'un SMA.

Dans la littérature on retrouve plusieurs définitions du concept « Agent » :

Selon Jacques Ferber : *"l'agent est une entité autonome physique ou virtuelle possédant des ressources propres qui est capable d'agir sur elle-même et sur son environnement, qui peut communiquer avec d'autres agents et dont le comportement est la conséquence de ses observations, de ses connaissances, et des interactions avec les autres agents [4]."*

S. Russell et P. Norvig le définissent ainsi : *"On appelle agent toute entité, qui peut être considérée comme percevant son environnement grâce à des capteurs et qui agit sur cet environnement via des effecteurs [5]."*

Selon L. Frécon et O. Kazar, un agent est une entité réutilisable ayant un accès contrôlé à des services et des ressources. Il peut également faire partie d'applications organisées en réseau d'agents collaborateurs [1].

Ces définitions mettent l'accent sur la caractéristique de l'autonomie permettant à l'agent de réaliser ses objectifs en exploitant efficacement les ressources dont il dispose.

Pour être distingué des autres entités logicielles, un agent possède plusieurs propriétés, les plus importantes sont [6] :

- **Son autonomie,** qui peut être définie en trois points :
  - un agent a sa propre existence, indépendante de l'existence des autres ;
  - un agent est capable de maintenir sa viabilité dans un environnement dynamique, sans avoir besoin d'un contrôleur externe ;
  - un agent peut prendre des décisions concernant son comportement futur (action, communication).
- **Sa proactivité,** un comportement proactif permet à l'agent d'interagir avec son environnement, et de prendre l'initiative au moment adéquat afin d'atteindre son objectif.

- **Sa sociabilité,** l'agent est en interaction avec d'autres agents pour les aider à réaliser leurs tâches.
- **Sa situation,** c'est la capacité d'agir et de modifier son environnement grâce à ses perceptions.
- **Sa mobilité,** un agent mobile peut se déplacer d'une machine à une autre pour exécuter des tâches de nature distribuée.
- **Son raisonnement et sa rationalité,** un agent est capable de raisonner rationnellement pour choisir les meilleures actions à entreprendre, afin d'optimiser sa productivité.

1.2.1.2 Types d'agents

Selon leurs modes de fonctionnement et la représentation de leurs environnements, les agents peuvent être classés en trois catégories, à savoir, les agents cognitifs, réactifs, et hybrides.

1. **Agent cognitif**

   Il a une représentation explicite et symbolique de son environnement et des autres entités du système et il est capable de raisonner. Ses principales caractéristiques sont la capacité de communiquer, de coopérer et de négocier. Il est capable de planifier son comportement pour atteindre ses objectifs. Son architecture est présentée dans la figure 1.1 [1].

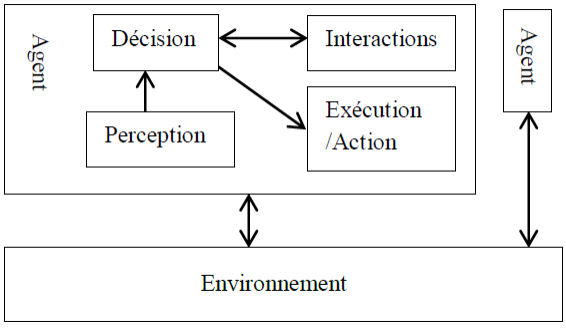


FIGURE 1.1. Architecture d'un agent cognitif

   Un agent cognitif possède les caractéristiques suivantes [4] :

   - **Sa croyance,** ensemble de connaissances que possède un agent sur lui-même, sur son environnement et sur les autres agents. Ces connaissances peuvent être des hypothèses.
   - **Son savoir,** ensemble de connaissances certaines (faits essentiels, relations structurelles et fonctionnelles ou bases de règles).

- **Son savoir-faire,** ensemble des compétences permettant de prendre des décisions et résoudre des problèmes.
- **Son contrôle,** ensemble d'intentions, de buts, de plans et de tâches.
- **Sa communication,** ensemble de protocoles de communication permettant à l'agent d'interagir avec d'autres agents.

2. **Agent réactif**

   Souvent qualifié de peu intelligent, un agent réactif est dirigé par des règles stimulus réponse et n'a pas de représentations explicites de son environnement [7]. Il peut agir sans nécessité de compréhension de son environnement, ni de ses objectifs. Un SMA composé d'agents réactifs possède communément un grand nombre (des milliers) d'agents. La figure 1.2 [1] illustre l'architecture de ce type d'agent.

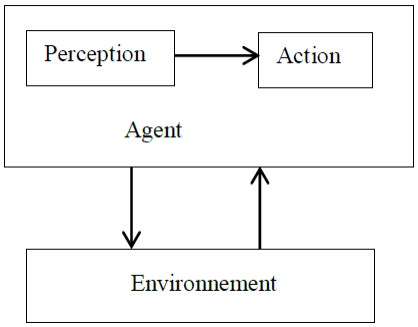


FIGURE 1.2. Architecture d'un agent réactif

3. **Agent hybride**

   Un agent hybride intègre les deux aspects : cognitif et réactif. L'idée consiste à combiner les deux types qui peuvent être vus comme complémentaires. Dans cette approche, l'agent est constitué de modules manipulant de façon indépendante les deux aspects.

1.2.2 Les systèmes multi-agents

Un SMA est considéré comme une population d'agents autonomes en interaction partageant un environnement commun.

1.2.2.1 Définition d'un SMA

Selon la définition de Jacques Ferber [4], Les SMA sont composés des éléments suivants :

1. **Un environnement E :** un espace qui dispose généralement d'une métrique.
2. **Un ensemble d'objets O :** qui peuvent être situés (pour tout objet, il est possible à un moment donné d'associer une position dans l'environnement E). Ils peuvent également être passifs (créés, détruits, manipulés et perçus par les agents).
3. **Un ensemble d'agents A :** ce sont les entités actives du système.
4. **Un ensemble de relations R :** pour unir les agents et les objets entre eux.
5. **Un ensemble d'opérations Op :** correspond aux différents types de manipulations (la perception, la production, la consommation, la transformation, etc.) appliquées par les agents A sur les objets O du système.
6. **Un ensemble d'opérateurs :** qui permettent de représenter l'application des opérations Op et la réaction du monde à cette modification (appelés : les lois de l'univers).

La figure 1.3 [4] illustre la notion de systèmes multi-agents.

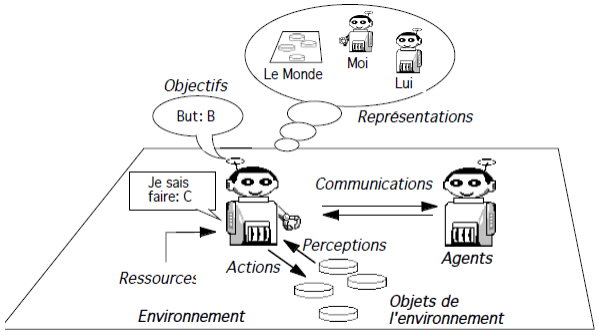


FIGURE 1.3. Système multi-agents

En résumé, un SMA est un ensemble d'agents autonomes partageant le même environnement et coopérant pour atteindre un objectif global ou des objectifs distincts.

1.2.2.2 Interactions dans un SMA

Jacques Ferber définit l'interaction comme : *« un ensemble de comportements résultant du regroupement d'agents qui doivent agir pour satisfaire leurs objectifs en tenant compte des contraintes*

provenant des ressources plus ou moins limitées, dont ils disposent et de leurs compétences individuelles [4].»

Il existe plusieurs modes d'interaction entre les agents :

- **La coopération**

  C'est la forme d'interaction la plus étudiée dans les SMA ; elle définit l'attitude sociale permettant d'augmenter les performances collectives du système [8]. Elle est basée, à la fois, sur la complémentarité d'intérêt et la confiance.

- **La coordination**

  La coordination permet d'articuler les actions accomplies par chacun des agents pour que l'ensemble des agents du système aboutisse à un objectif général cohérent et performant. Elle permet d'assurer la coopération entre des agents autonomes [9].

  La coordination est primordiale pour quatre principales raisons :

  1. Les agents ont besoin d'informations et de résultats fournis par d'autres agents ;
  2. Les ressources sont limitées ;
  3. On cherche à optimiser les coûts ;
  4. On veut permettre aux agents ayant des objectifs distincts, mais dépendants les uns des autres de satisfaire ces objectifs et d'accomplir leur travail en tirant éventuellement parti de cette dépendance.

- **La négociation**

  La négociation sert à améliorer les accords (en réduisant les inconsistances et l'incertitude) sur des points de vue communs ou des plans d'action, grâce à l'échange structuré d'informations pertinentes. Son protocole minimal d'actions peut être résumé comme suit : (proposer, évaluer, accepter ou refuser une solution).

1.2.2.3 La communication dans un système multi-agents

La communication dans les SMA, comme chez les humains, est la base des interactions et de l'organisation sociale. Elle peut être sélective ou diffusée.

Il existe principalement deux modes de communication :

1. **Communication par envoi de messages**

   Dans les systèmes fonctionnant avec ce type de communication, les connaissances sont distribuées entre les différents agents. Chacun d'eux communique directement avec les autres en mode point à point ou par diffusion.

2. **Communication par partage d'informations**

   Dans les systèmes où la communication se fait par partage de ressources, il n'y a pas une liaison directe entre les différents agents. Ils partagent une zone de données commune appelée tableau noir (Blackboard). Le partage d'informations est utilisé quand il y a recouvrement des domaines d'expertise de chaque agent et quand ils possèdent une connaissance limitée sur les autres agents, d'où des problèmes de synchronisation sont posés [1].

1.2.2.4 L'apprentissage chez l'agent

L'autonomie décisionnelle est l'une des caractéristiques les plus importantes chez un agent, lui permettant de prendre des décisions de façon indépendante. Grâce à l'apprentissage, l'agent est capable d'apprendre, de s'adapter et d'améliorer sa performance individuelle et collective au sein du système.

1. **Apprentissage mono agent (centralisé)**

   L'apprentissage est dit mono agent, lorsque le processus d'apprentissage est exécuté dans toutes ses parties par un seul agent et ne requiert aucune interaction avec d'autres agents. Dans ce type d'apprentissage, les perceptions d'un agent servent à choisir des actions pour améliorer ses compétences individuelles, évoluer, s'adapter et l'aider à agir dans le futur. Face à une situation nouvelle, l'agent prend beaucoup de temps pour percevoir son environnement. Avec un module d'apprentissage, la performance et la rapidité d'un agent augmentent avec l'exécution de tâches similaires. Son comportement, dans ce cas, passe d'un état délibératif à un état réactif. Dans un système multi-agents, un seul agent peut participer simultanément dans plusieurs processus d'apprentissage. De plus, plusieurs apprenants centralisés tentent d'atteindre des objectifs d'apprentissage distincts ou identiques et ils peuvent être actifs en même temps [10].

2. **Apprentissage multi-agent (décentralisé)**

   L'apprentissage est dit multi-agent si plusieurs agents sont engagés dans le même processus d'apprentissage. Les agents, dans ce cas, apprennent d'une manière interactive et décentralisée comme une entité cohérente [11].

   Dans un environnement multi agent, les agents ont la possibilité d'apprendre grâce aux autres ; ils peuvent également apprendre à propos des autres agents. Grâce à ces perceptions, l'agent peut construire un modèle de comportement de l'autre agent. Ce modèle lui permet de prédire les actions futures de l'autre agent, pour assurer une coordination ou une collaboration pour réaliser un objectif commun.

1.2.2.5 Application des SMA à la médecine

Les systèmes dédiés à la santé font partie des systèmes les plus compliqués à aborder, vu les enjeux qu'ils rencontrent (la complexité, la disponibilité des données, la méfiance des patients, de la société et des médecins envers ces structures, les produits de santé et les logiciels,etc.). Les techniques classiques de l'IA restent limitées pour modéliser des environnements avec une telle variété d'utilisateurs, ainsi que la complexité des processus et des interactions. L'application du paradigme « Agent » et « SMA » dans le domaine médical permet d'améliorer la performance des systèmes informatiques en termes de facilité de maintenance, de réutilisation, de portabilité, de fiabilité, de robustesse, d'évolutivité, de flexibilité et de réduction des coûts. Les SMA sont mis à la disposition du personnel médical pour faciliter la prise de décision, traiter des données provenant de sources différentes et sous des formats différents, améliorer la qualité des soins, des traitements à prescrire aux patients et réduire les charges financières et le temps de traitement concernant les prestations de service informatique.

On retrouve le paradigme « Agent » dans une large gamme d'applications médicales :

- Des systèmes d'aide à la décision ;
- Gestion des données médicales et l'accès aux sources de données distribuées ;
- Télémédecine ;
- Traitement d'images médicales et simulation ;
- Planification et allocation des ressources.

1.3 Apprentissage automatique et apprentissage profond

L'apprentissage automatique ou Machine Learning (ML) est une branche de l'IA qui consiste à donner aux machines l'apptitude d'apprendre et d'améliorer leurs performances automatiquement en fonction des données qu'ils traitent sans être explicitement programmés [8]. L'apprentissage automatique a pour objectif de définir une fonction f : X→ Y reliant les entrées X et les sorties Y, et dépendant du type d'algorithme d'apprentissage utilisé [12].

1.3.1 Catégories de ML

Le type et le volume des données utilisées pour la construction des algorithmes de l'apprentissage automatique permettent de définir les différents types de ce dernier [13]. On peut distinguer trois catégories principales d'apprentissage : supervisé, non supervisé et par renforcement.

1.3.1.1 Apprentissage supervisé

Dans cette approche, l'algorithme prend en entrée un ensemble de données d'apprentissage bien défini, incluant les solutions souhaitées (étiquettes) [14]. L'apprentissage supervisé a pour but d'identifier une fonction déterministe qui attribue une entrée à toute sortie et de l'appliquer sur un processus analytique pour la prédiction de futures observations en réduisant le taux d'erreur. Les algorithmes sont entraînés en utilisant un sous-ensemble d'exemples prétraités et les performances des algorithmes sont évaluées à l'aide d'un sous-ensemble de test.

On peut distinguer, selon le type des étiquettes, deux types de problèmes d'apprentissage supervisé : classification et régression (figures 1.4).

- **Classification :** le processus de classification consiste à déterminer le modèle ou la fonction qui permet de séparer les données en plusieurs classes catégorielles (valeurs discrètes).
- **Régression :** le processus de régression consiste à trouver le modèle ou la fonction permettant de distinguer les données en valeurs réelles continues ou identifier le mouvement de distribution en fonction des données historiques.

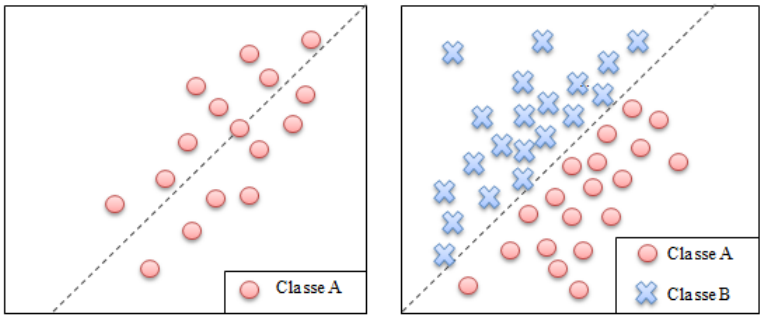


FIGURE 1.4. Apprentissage supervisé

Il existe plusieurs algorithmes d'apprentissage supervisé, les plus connus sont [14] :

- **K-NN (K nearestneighbors) ou les K plus proches voisins :**peut-être défini comme un algorithme non paramétrique, il calcule la distance entre les données de test et l'entrée et donne la prédiction appropriée.
- **La régression linéaire :** c'est une méthode statistique qui consiste à trouver la droite la mieux ajustée passant par les points.
- **La régression logistique :** une méthode statistique appliquée pour analyser un ensemble de données avec une ou plusieurs variables indépendantes afin de déterminer un résultat. La méthode vise à trouver le meilleur modèle qui définit la relation entre les entrées et les sorties. Elle est utilisée pour la classification et la régression.
- **SVM (Support Vector Machine) ou machine à vecteur de support :** c'est une méthode de classification non linéaire et non paramétrique. Son objectif principal est de trouver un hyperplan pour la séparation des données. Elle est efficace en mémoire pour les espaces de grande dimension.
- **Arbre de décision ou decision trees (DTs) :** ce sont des structures de type organigramme, où chaque nœud interne désigne un test sur un attribut, chaque branche représente un résultat du test et chaque nœud terminal contient une étiquette de classe.
- **La forêt aléatoire ou random forest (RF) :** elle peut être définie comme un ensemble de DTs, où chaque arbre individuel construit une classe de sortie, puis la moyenne des prédictions est calculée. Le résultat final est généré en prenant le mode des classes trouvées séparément.
- **Le réseau de neurones artificiels ou artificial neural network (ANN) :** un modèle basé

sur le raisonnement humain. Son architecture est une hiérarchie de couches (couches d'entrée, couches cachées et couches de sortie). Les données sont reçues par la couche d'entrée, puis transmises à une couche cachée pour le traitement et la fourniture des résultats d'apprentissage à la couche de sortie. La couche de sortie affiche les résultats de la classification.

1.3.1.2 Apprentissage non supervisé

Il est mieux adapté aux problèmes nécessitant de grandes masses de données non étiquetées. Les algorithmes, dans ce cas, segmentent les données en ensembles d'exemples (clusters) ou en groupes de fonctionnalités. L'apprentissage non supervisé est un processus itératif d'analyse de données et de création de modèles sans intervention humaine. L'objectif de l'apprentissage non supervisé est d'analyser les données d'entrée et de réduire leur dimensionnalité. Il existe trois types de problèmes d'apprentissage non supervisé : clustering, association et réduction de dimensionnalité [14].

- **Clustering :** c'est un moyen permettant de regrouper des données non étiquetées présentant des propriétés similaires dans différents clusters (figure 1.5).
  Cette technique est utile pour la compression, la segmentation d'images,etc.
- **Association :** une technique qui permet de déterminer les relations entre les variables d'un ensemble de données en appliquant différentes règles.
- **Réduction de la dimensionnalité :** cette technique est privilégiée lorsque le nombre de caractéristiques dans un ensemble de données est trop élevé. Elle permet de réduire la taille de l'ensemble de données d'entrée à une taille gérable, tout en gardant les éléments importants de ce dernier.

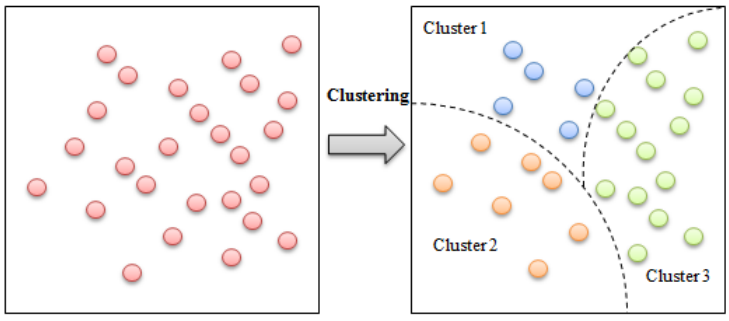


FIGURE 1.5. Apprentissage non supervisé (Clustering)

Les algorithmes d'apprentissage non supervisé les plus courants sont :

- **K-moyennes (k-means) :** c'est un algorithme de clustering permettant de partitionner les données similaires en un ensemble de k clusters mutuellement exclusifs, où k est prédéfini [15]. Cet algorithme est évolutif et offre la possibilité de traiter de grandes masses de données. Cependant, il peut converger vers un minimum local.
- **Le clustering hiérarchique (HCA) :** une approche alternative à l'algorithme k-moyennes pour identifier les clusters dans un ensemble de données. Il crée une hiérarchie de clusters sans la spécification de leur nombre.
- **L'analyse en composantes principales ou principal component analysis (PCA) :** une technique qui réduit la dimensionnalité d'un ensemble de données, augmente l'interprétabilité et minimise la perte d'informations. Ceci est assuré par la création de nouvelles variables non corrélées, qui maximisent successivement la variance [16].
- **Algorithme Apriori :** un algorithme de recherche de règles d'association conçu pour l'exploration des données. Il sert à identifier les propriétés les plus fréquentes dans un ensemble de données et d'en déduire une catégorisation.

1.3.1.3 Apprentissage par renforcement

Dans ce type d'apprentissage, le système est composé d'agents ayant la capacité d'observer l'environnement, d'effectuer des actions et d'obtenir des récompenses ou des pénalités en retour [14]. Il est capable de déterminer, de façon autonome, la meilleure stratégie (politique) pour maximiser le nombre de récompenses au fil du temps.

1.3.2 Apprentissage profond

L'apprentissage profond est une sous-catégorie de l'apprentissage automatique. Son principe est inspiré de l'anatomie du cerveau humain, il est basé sur des réseaux de neurones artificiels. Le processus d'apprentissage est qualifié de profond parce que la structure des réseaux neuronaux artificiels se compose de plusieurs couches d'entrée, de sortie et intermédiaires. Chaque couche contient des unités qui transforment les données d'entrée en informations. Ces informations peuvent être utilisées par la couche suivante pour une tâche prédictive spécifique. Grâce à cette structure, la machine est capable de raisonner, d'apprendre, de représenter des connaissances

d'une façon structurée et de planifier des tâches.

1.3.3 Processus de l'apprentissage automatique

L'apprentissage automatique offre l'opportunité d'anticiper les changements et de prédire le futur, en appliquant certaines stratégies sur un ensemble de données. La précision du modèle généré par le processus du ML, dépend de la qualité et de la quantité des données d'entraînement qui lui sont intégrées. Dans cette section, nous allons détailler le processus du ML (voir figure 1.6).

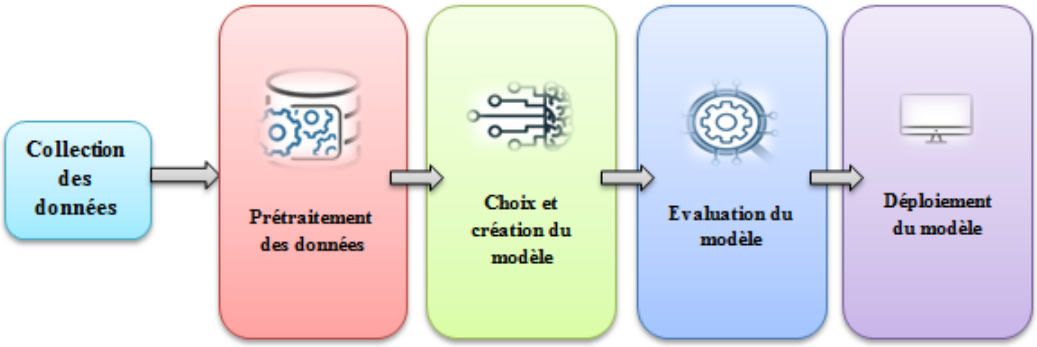


FIGURE 1.6. Processus de l'apprentissage automatique

1.3.3.1 Prétraitement des données

Dans le processus de l'apprentissage automatique, la phase du prétraitement des données est une phase importante. Elle permet d'obtenir des résultats de prédiction plus précis. Pour rendre les données d'entrée exploitables par les algorithmes d'apprentissage automatique, cette phase comprend généralement les étapes suivantes en fonction de la source et du format des données :

1. **Interpolation des valeurs manquantes :** Le problème des données manquantes est un problème courant. Il peut survenir pour de nombreuses raisons : erreurs humaines, dommages matériels, mesures imprécises ou perdues, etc. Cela peut conduire à un ralentissement du traitement analytique et à des conclusions erronées. Pour faire face à ce problème, de nombreuses stratégies ont été développées :
   (a) ***Suppression des données manquantes :*** elle consiste à supprimer tous les cas avec des valeurs manquantes. Elle est privilégiée pour les systèmes, où les données manquantes ne sont pas pertinentes et n'affectent pas les résultats obtenus.

(b) ***Imputation :*** elle consiste à remplacer les valeurs manquantes par d'autres valeurs (moyenne, médiane, constante,etc.).

(c) ***Imputation multiple :*** dans cette méthode, on se base sur l'incertitude concernant les données manquantes en créant plusieurs ensembles de valeurs plausibles et en combinant de manière appropriée les résultats obtenus à partir de chacun d'eux.

(d) ***Modélisation prédictive :*** elle permet de créer un modèle prédictif qui sert à estimer les valeurs qui remplaceront les valeurs manquantes.

2. **Synchronisation des données :** elle garantit la sécurité, la cohérence et la précision des données et assure l'harmonie entre les différentes sources de données. Pour éviter les erreurs et prévenir les atteintes à la vie privée, les modifications sont mises à niveau dans chaque système en temps réel.

3. **Mise à l'échelle (normalisation des données) :** durant cette étape, les variables indépendantes de l'ensemble de données sont standardisées en utilisant une échelle commune sans fausser les différences dans les plages de valeurs, ou perdre des informations. Un processus de normalisation peut également être appliqué pour convertir les valeurs numériques d'origine en valeurs nominales pour un algorithme spécifique [17]. Le processus de normalisation est appliqué généralement pour deux principales raisons : apprentissage rapide et réduction de la complexité du problème. Dans la littérature, plusieurs méthodes de normalisation des données ont été proposées [18] :

   (a) ***Normalisation Min-Max***

   Cette technique consiste à mapper les données d'entrée dans une plage prédéfinie [0,1] ou [-1,1]. La méthode min-max normalise les valeurs des attributs X d'un ensemble de données en fonction de ses valeurs minimales et maximales. La valeur x_m de l'attribut X est convertie en x'_m dans la plage [*low, high*] en appliquant l'équation :

$$x'_m = low + \frac{(high - low)(x_m - minZ)}{maxZ - minZ} \tag{1.1}$$

   Le principal problème de l'utilisation de la méthode de normalisation min-max dans les prévisions de séries chronologiques est que les valeurs minimales et maximales de l'ensemble de données hors échantillon sont inconnues.

(b) ***Normalisation Z-score***

Cette technique est basée sur la moyenne et l'écart type des valeurs des attributs X. Les nouvelles valeurs des attributs sont calculées selon l'équation suivante :

$$x'_m = \frac{x_m - \mu_m}{\delta_m} \tag{1.2}$$

Où μ_m est la valeur moyenne de l'attribut, δ_m est la dérivation standard, x_m sont les données d'entrée et x'_m représente la donnée après la normalisation.

L'inconvénient de cette méthode est qu'elle ne garantit pas un intervalle commun pour les scores normalisés provenant de différents systèmes.

(c) ***Normalisation par la médiane***

La valeur normalisée x'_m est obtenue en calculant la médiane des valeurs des données d'entrée x_m de l'attribut X selon l'équation :

$$x'_m = \frac{x_m}{mediane(x_m)} \tag{1.3}$$

Elle est efficace lorsqu'il est nécessaire de calculer le rapport entre deux échantillons hybrides.

(d) ***Normalisation par la méthode tangente hyperbolique "Tanh"***

Les données sont normalisées par l'équation suivante :

$$\frac{1}{2}[\tanh(0.001\frac{x_m - \mu}{\delta} - 1)] \tag{1.4}$$

Où : μ est la moyenne arithmétique, δ l'écart-type des données et $\tanh$ est la tangente hyperbolique. Cette méthode met chaque score normalisé dans l'intervalle [0, 1].

4. **Augmentation des données :** pour améliorer la performance du modèle proposé et lui donner plus d'informations avec lesquelles travailler, parfois il est nécessaire d'enrichir les données existantes par de nouvelles données externes. Cette phase est appliquée si l'ensemble de données est insuffisant pour l'apprentissage et le test. Son objectif principal est d'éviter les problèmes de sur-apprentissage (overfitting) et de sous-apprentissage (underfitting).
5. **Extraction et sélection des caractéristiques :** vise à extraire et déduire des informations

de l'ensemble d'entités d'origine pour créer un nouveau sous-espace d'entités. Son idée principale est de compresser les données pour conserver les informations importantes [19]. La méthode d'extraction de caractéristiques la plus populaire est l'analyse en composantes principales (PCA). C'est une méthode non paramétrique utilisée pour l'extraction d'informations pertinentes dans un ensemble de données souvent redondantes ou bruitées. Dans la littérature, de nombreuses variantes de PCA ont été proposées.

Il existe une autre approche pour la réduction de la dimensionnalité d'un ensemble de données qui est la sélection des caractéristiques. Elle permet de réduire le volume des données d'entrée en sélectionnant un sous-ensemble de caractéristiques les plus pertinentes parmi les caractéristiques d'origine. L'objectif principal de cette approche est d'améliorer l'efficacité de calcul du modèle, de réduire sa complexité et de minimiser l'erreur de généralisation introduite causé par les données bruitées, redondantes et non pertinentes [20]. Le processus de sélection des caractéristiques est illustré dans la figure 1.7 [21].

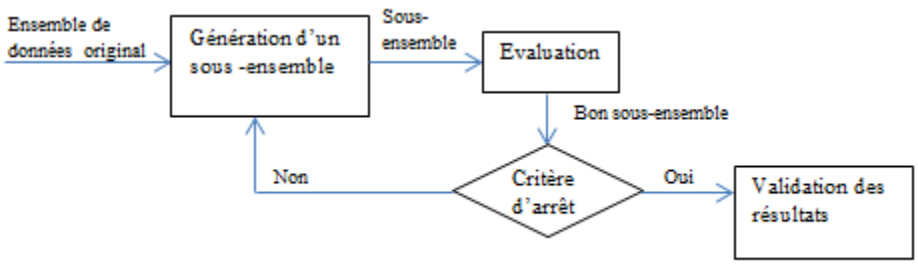


FIGURE 1.7. Processus de sélection des caractéristiques

Il existe plusieurs approches de sélection de caractéristiques : (***approche de filtrage***, ***approche Wrapper*** et ***approche embarquées***). Dans le tableau 1.1 [22], une comparaison entre les différentes approches de sélection des caractéristiques est présentée.

TABLE 1.1. Tableau comparatif entre les approches de sélection des caractéristiques

Approche / Critère	Approche de filtrage	Approche Wrapper	Approches embarquées
Technique	Mesures statistiques	Algorithme d'optimisation	Combinaison des Approche de filtrage et approche Warpper
Temps de calcul	Gain de temps	lente	Lente
Coût de calcul	Coût Réduit	Coûteuse	Coûteuse
Espace de calcul	Espace réduit	Coûteuse	Coûteuse
Complexité	faible complexité	Complexité élevée	Complexité élevée

Le but de la sélection et l'extraction des caractéristiques est de réduire le nombre de caractéristiques de l'ensemble de caractéristiques d'origine, afin de réduire la complexité et le sur-ajustement du modèle, améliorer l'efficacité du calcul du modèle et réduire l'erreur de généralisation.

1.3.3.2 Choix du modèle

Durant cette phase, l'algorithme ou les algorithmes appropriés, sont choisis selon l'objectif et le type des données d'entrée. Par exemple, certains modèles sont plus adaptés pour traiter des textes, tandis que d'autres modèles sont adaptés pour traiter des images. Ensuite, l'algorithme choisi peut être entraîné, testé et validé après avoir divisé l'ensemble d'entrée en trois sous-ensembles (sous-ensemble d'entraînement, sous-ensemble de test et sous-ensemble de validation).

1.3.3.3 Evaluation du modèle

C'est une étape très importante dans le processus de ML. Elle sert à évaluer le modèle proposé, en utilisant différentes métriques qui dépendent entièrement du type et du plan de mise en œuvre du modèle d'apprentissage. Dans cette section, nous allons détailler les métriques d'évaluation utilisées pour les modèles de classification [23].

1. **La matrice de confusion**

   La matrice de confusion est un outil de mesure de performance qui donne des informations sur les cas réels et prédits (voir tableau 1.2 [13]). Il permet de savoir à quel point le modèle est confus.

TABLE 1.2. Matrice de confusion

REEL	PREDICTION	
	Positif	Négatif
Positif	Vrai positif (TP)	Faux négatif (FN)
Négatif	Faux positif (FP)	Vrai négatif (TN)

Où :

Vrai positif (TP) : les valeurs réelles et prédites, sont identiques et positives.

Vrai négatif (TN) : les valeurs réelles et prédites, sont identiques et négatives.

Faux positif (FP) : les valeurs réelles et prédites sont différentes. Valeurs négatives incorrectement identifiées comme positives.

Faux négatif (FN) : les valeurs réelles et prédites sont différentes. Valeurs positives incorrectement identifiées comme négatives.

2. **L'exactitude (Accuracy)**

   C'est la proportion du nombre total de prédictions correctes. Sa valeur est comprise entre 0 et 1 pour une mauvaise précision allant jusqu'à une précision parfaite.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.5}$$

3. **La précision (PPV/ precision)**

   C'est la proportion de vrais positifs sur tous les résultats positifs. Le résultat est compris entre 0 et 1 et est calculé par l'équation suivante :

$$PPV = \frac{TP}{TP + FP} \tag{1.6}$$

4. **Sensibilité (TPR/ recall)**

   C'est la fraction de valeurs positives qui sont classées correctement. Sa valeur est comprise entre 0 et 1.

$$TPR = \frac{TP}{TP + FN} \tag{1.7}$$

5. **Spécificité (specificity)**

   La proportion d'exemples négatifs, qui ont été prédits comme positifs et pourraient être qualifiés de faux positifs.

$$Specificity = \frac{TN}{TN + FP} \tag{1.8}$$

6. **FPR (False Psitive Rate)**

   le pourcentage de cas négatifs qui ont été incorrectement prédits comme positifs.

$$FPR = \frac{FP}{FP + TN} \tag{1.9}$$

7. **Score F1**

   C'est la moyenne harmonique entre la sensibilité et la valeur prédictive positive. Un score F1 élevé signifie que les deux mesures sont élevées.

$$Score-F1 = \frac{2 * (PPV * TPR)}{PPV + TPR} \tag{1.10}$$

8. **La courbe ROC**

   Dite fonction d'efficacité du récepteur, c'est une mesure de performance graphique utilisée pour les classificateurs binaires. Elle est représentée sous forme d'une courbe, qui donne le taux de vrai positif (sensibilité) en fonction du taux de faux positifs (1 – spécificité).

9. **La zone sous la courbe ROC (AUC)**

   L'AUC d'un classificateur représente la probabilité que la valeur d'une instance positive sélectionnée au hasard, soit plus élevée que la valeur d'un exemple négatif sélectionné au hasard. Cette métrique est calculée pour évaluer les performances du classificateur et permet de fournir un examen de la stabilité et de la cohérence de ce dernier.

1.3.4 Apprentissage automatique appliqué à la médecine

Dans le domaine de la santé, l'application la plus courante du ML est la médecine de précision [24]. Par exemple, prédire le protocole de traitement susceptible de réussir sur un patient en fonction

de plusieurs attributs. Il peut être appliqué dans les quatre phases du processus médical : prévention, détection, diagnostic et traitement (Figure 1.8) pour :

- **Identification des maladies et diagnostic :** En appliquant les approches de l'apprentissage automatique, il est possible de détecter les maladies à leur début. Par conséquent, cela pourra aider les médecins à gagner du temps, à minimiser les coûts et la complexité.
- **Découverte et production de médicaments :** ML peut être utilisé pour la découverte, la production et la personnalisation des médicaments selon les pathologies, en réduisant les coûts et le temps du processus de fabrication.
- **Chirurgie assistée par robots :** l'apprentissage automatique est également utile en chirurgie, il peut aider à identifier différentes parties du corps et même à effectuer une intervention chirurgicale.
- **Analyse des données médicales :** cette tâche peut être très efficace pour mieux comprendre les causes de l'évolution des maladies, les prédire et même les prévenir. L'application du ML peut aider les médecins à interpréter les données d'une manière optimisée et à sauver la vie des patients, en leur faisant gagner du temps.

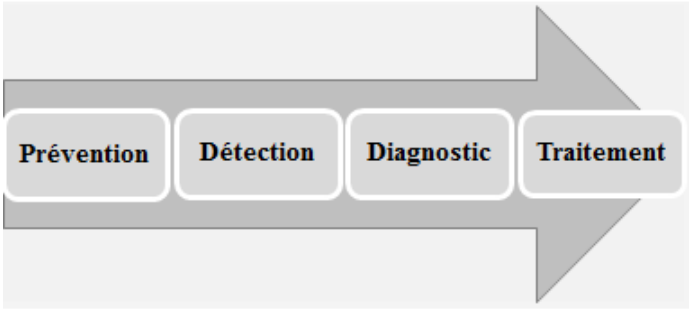


FIGURE 1.8. Phases du processus médical

1.4 Métaheuristiques

Les métaheuristiques sont largement utilisées pour résoudre des problèmes d'optimisation difficiles. Dans cette section, nous allons proposer un aperçu sur les métaheuristiques et les principales méthodes d'optimisation.

1.4.1 Définition d'une métaheuristique

Une méthode approximative, dite « heuristique », est une méthode de résolution algorithmique permettant de fournir rapidement des solutions réalisables à un problème décisionnel en temps polynomial. Elle est conçue pour un problème particulier et s'appuie sur sa propre structure pour éviter l'explosion combinatoire, n'explorant qu'une partie de l'espace des combinaisons. Des heuristiques générales et plus poussées, qui s'appliquent à différents problèmes donnent naissance à une nouvelle famille de méthodes approximative : métaheuristiques. Les heuristiques et métaheuristiques font partie des méthodes stochastiques où l'algorithme donne différentes solutions entre deux exécutions indépendantes, à cause de l'aspect aléatoire.
Les métaheuristiques sont des méthodes génériques, non-déterministes, souvent inspirées du processus naturels et sont hybridées avec d'autres méthodes de recherche opérationnelle. Elles permettent d'optimiser une large gamme de problèmes différents, sans avoir besoin de changements profonds dans l'algorithme utilisé. On peut classer les métaheuristiques, en se basant sur le type de solution en deux classes distinctes :

— **Métaheuristiques à base de solution unique :** manipulant un seul point à la fois, en la faisant évoluer sur l'espace de recherche à chaque itération.
— **Métaheuristiques à base de population :** manipulant un ensemble de solutions en parallèle, à chaque itération pour une meilleure discrimination de l'espace de recherche.

1.4.1.1 Métaheuristiques à base de solution unique

Ce sont des méthodes de recherche locale ou méthodes de trajectoire, leur mécanisme s'appuie sur l'évolution itérative d'une solution dans l'espace de recherche pour l'obtention d'un optimum global [25]. Les méthodes les plus répandues dans la littérature sont : le recuit simulé, la méthode de descente, la recherche tabou et la méthode GRASP.

1. ***Le recuit simulé (SA : simulated annealing)***

   C'est une méthode inspirée de la métallurgie. Elle tire ses origines des expériences effectuées par Metropolis et al. en 1953 [26]. Elle a été mise en oeuvre par trois chercheurs de la société IBM en 1983 [27], et indépendamment par Cerny en 1985 [28]. Son principe, inspiré du recuit physique, consiste à faire des séries de refroidissement lent et de réchauffage

d'un matériau pour minimiser son énergie. La simulation de recuit peut être utilisée pour trouver une approximation d'un minimum global pour une fonction à plusieurs variables. La notion de refroidissement lent est interprétée, comme une diminution lente, de la probabilité d'accepter des solutions pires, à mesure que l'espace de la solution est exploré, ceci permet une recherche plus approfondie de la solution optimale globale. On peut appliquer l'algorithme du recuit simulé sur plusieurs problèmes d'optimisation, tels que la définition de l'enchainement des tâches, dans un processus de fabrication, etc.

2. ***La recherche tabou (TS : Tabu Search)***

   TS est introduite par Fred Glover en 1986 [29]. Son principe est inspiré du fonctionnement de la mémoire humaine, elle utilise une mémoire (basé sur l'historique de recherche) appelée liste tabou pour enregistrer les dernières solutions rencontrées ou des caractéristiques de solutions, vers lesquelles il est interdit de se déplacer pour éviter le problème de l'optimum local.

3. ***Méthode de descente***

   La méthode de descente fait partie des méthodes les plus intuitives et les plus simples. Son fonctionnement consiste à choisir, à chaque itération, un point dans le voisinage de la solution courante à partir d'une solution initiale pour améliorer la fonction objective. L'algorithme s'arrête lorsque l'amélioration du voisinage devient impossible. Il existe différentes stratégies pour le choix du voisinage ; l'algorithme Hill Climbing consiste à choisir celle avec la meilleure fitness par rapport à toutes les autres solutions. Le First Improvement Hill Climbing vise à choisir le premier voisin améliorant rencontré.

4. ***La méthode GRASP***

   La procédure de recherche gloutonne aléatoire adaptative (GRASP : Greedy Randomized Adaptive Search Procedure) est proposée par Feo et Resende [30] [31]. C'est une métaheuristique à départs multiples, itérative, dépourvue de mémoire. Chaque itération consiste de deux phases principales : une phase de construction, pour générer une solution réalisable et une phase de recherche locale, pour trouver un optimum dans le voisinage de l'élément construit. L'algorithme GRASP se termine et la meilleure solution trouvée est conservée après un nombre donné d'itérations.

1.4.1.2 Métaheuristiques à base de population

Les métaheuristiques à base de population de solutions partent d'un ensemble de solution, contrairement aux métaheuristoiques à base de solution unique. Cela permet d'améliorer, au fil des successions des itérations, toute une population de solutions. La population, dans ces méthodes est utilisée comme facteur de diversité. Dans cette classe, nous pouvons distinguer deux grandes catégories :

— Les algorithmes évolutionnaire.

— Intelligence en essaims.

1. ***Les algorithmes évolutionnaires***

   Les algorithmes de cette catégorie s'inspirent de la théorie d'évolution naturelle (bio inspirés) pour la résolution de problèmes complexes. Ce sont des algorithmes itératifs, appliquant des opérateurs stochastiques sur un ensemble d'individus. Tout individu soumis à l'évolution et appartenant à l'espace de recherche du problème d'optimisation est considéré comme solution provisoire. Durant la phase initiale d'un algorithme évolutionnaire, la population est générée de façon aléatoire et elle évolue itérativement, jusqu'à atteindre un critère d'arrêt pour concevoir les générations de l'algorithme, en appliquant une succession d'opérateurs, à savoir, un opérateur de sélection, un opérateur de croisement et un opérateur de mutation permettant d'engendrer la nouvelle population à la génération suivante [32]. Le principe des algorithmes évolutionnaires est présenté dans la figure 1.9 [32].

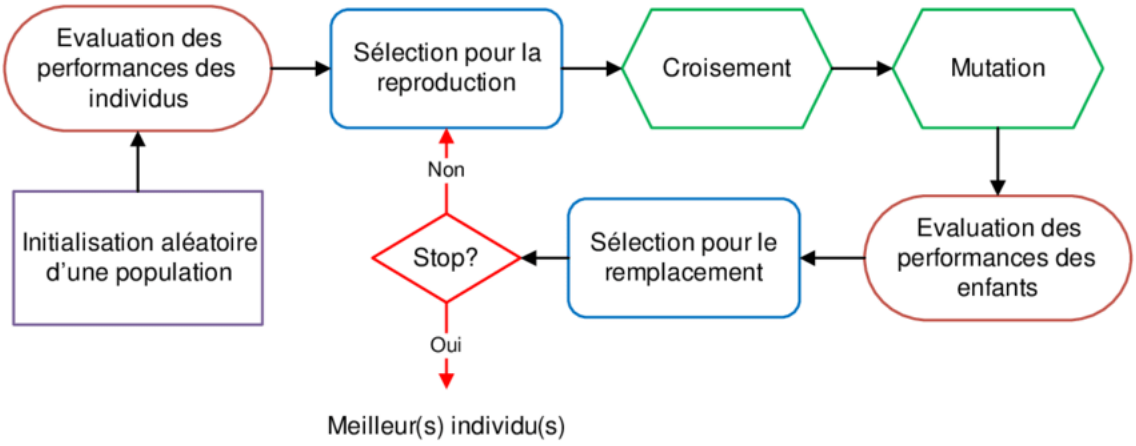


FIGURE 1.9. Principe de l'algorithme évolutionnaire

Il existe plusieurs sous-classes d'algorithmes évolutionnaires, dans cette section nous allons

présenter deux principaux algorithmes évolutionnaires (algorithmes génétiques et Harmony Search) :

(a) **Algorithmes génétiques**

Les algorithmes génétiques (AG) ont été conçus dans les années 1970 par John Holland et David Goldberg [33] [34]. Ils tirent leur principe des mécanismes biologiques (lois de Mendel, principe fondamental de Charles Darwin). Ils simulent le processus d'évolution d'une population en s'appuyant sur cinq phases principales (figure 1.10 [35]) :

i. **Génération de la population initiale :** Le processus commence avec un ensemble de solutions initiales appelé population. Chaque individu est défini par un ensemble de gènes, qui forment un chromosome (solution).

ii. **Fonction de fitness :**La fonction de fitness détermine la capacité d'un individu à rivaliser avec d'autres individus (condition physique). Elle permet également de donner un score de fitness à chaque individu. Ce score détermine la probabilité la sélection d'un individu pour la reproduction.

iii. **Sélection :** L'intérêt à travers cette phase est de sélectionner les individus les plus adéquats et de les laisser transmettre leurs gènes à la génération suivante. Deux paires d'individus (parents) sont sélectionnés en fonction de leurs scores de fitness. Les individus ayant une bonne condition physique ont plus de chances d'être sélectionnés pour la reproduction.

iv. **Croisement :** C'est une étape cruciale dans un AG. Elle permet de combiner deux individus parents pour générer des individus enfants tout en conservant les bonnes caractéristiques des parents.

v. **Mutation :** La mutation a pour but la diversification des populations en appliquant des modifications aléatoires sur les gènes d'un individu sélectionné, dans le but d'éviter la convergence vers des solutions optimales.

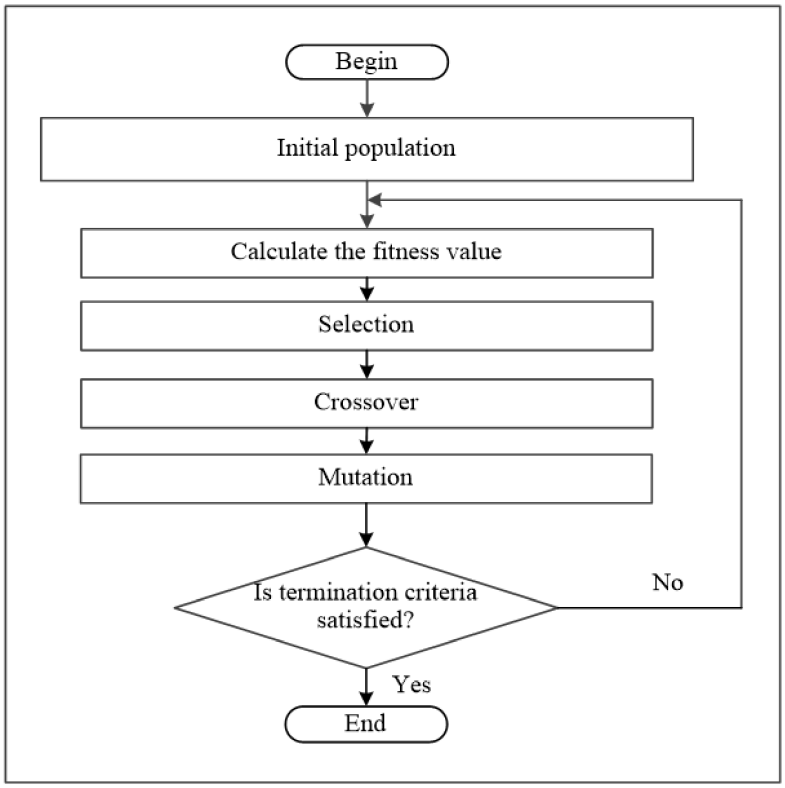


FIGURE 1.10. Processus d'un algorithme génétique

(b) **Harmony search**

Proposé en 2001[36], l'algorithme HS imite le processus de l'improvisation musicale, où chaque musicien improvise, avec son instrument, des sons pour créer une bonne harmonie (état parfait d'harmonie). La figure 1.11 montre l'analogie entre l'improvisation musicale et l'optimisation, où un musicien correspond à une variable de décision, le registre de l'instrument de musique correspond à la plage de chaque valeur de variable, l'harmonie musicale à un instant donné correspond au vecteur solution à une itération donnée et l'esthétique musicale correspond à la fonction objectif. Comme la qualité de l'haromie musicale qui s'améliore répétition après répétition, le processus d'optimisation améliore la solution courante au fil des itérations.

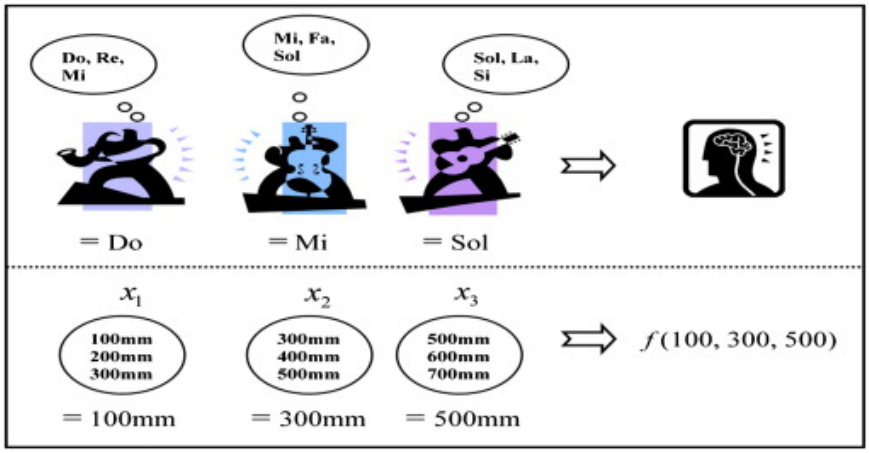


FIGURE 1.11. Analogie entre improvisation musicale et optimisation

Le processus de l'algorithme HS pour l'optimisation consiste en quatre principales étapes [37] :

i. **Initialisation de la mémoire (HM : harmony memory) :** Initialement, la HM est constituée d'un certain nombre de solutions sous forme de vecteurs générées aléatoirement.

ii. **Improvisation d'une nouvelle harmonie :** Chaque composant de cette solution est obtenu sur la base de HMCR (en anglais : Harmony Memory Considering Rate). Le HMCR est défini comme la probabilité de sélectionner un composant, parmi les membres du HM, et 1-HMCR est donc la probabilité de le générer de manière aléatoire.

iii. **Mise à jour de la HM :**La nouvelle solution de l'étape 2 est évaluée et comparée avec la plus mauvaise de toutes les valeurs des solutions insérées dans HM. Si elle est meilleure en termes de fonction objectif, elle sera insérée dans HM et remplacera l'ancienne valeur de la plus mauvaise solution. Sinon, elle sera rejetée.

iv. **Critère d'arrêt :** Le processus d'itérations dans les étapes 2 et 3 se termine lorsqu'un nombre prédéfini d'improvisations est atteint. Le meilleur vecteur dans la HM est sélectionné et est considéré comme solution optimale au problème étudié.

2. ***Intelligence en essaim***

Intelligence en essaim ou (SI : Swarm Intelligence en anglais) fait référence à la modélisation mathématique et informatique des phénomènes biologiques. Elle consiste à fournir des

groupes d'individus (agent) artificiels, dont la capacité individuelle est simple et très limitée, mais les interactions locales entre eux et avec leur environnement permettent de réaliser des tâches complexes, nécessaires à leur survie. La structure de chaque agent, fait appel à une représentation et un mécanisme de raisonnement basique et simple. Le comportement collectif global auto-organisé émerge des interactions locales entre ces agents. Intelligence en essaim recouvre de nombreux algorithmes, tels que l'optimisation des colonies de fourmis, l'optimisation des essaims de particules (PSO : practicle swarm optimization), les colonies d'abeilles et les systèmes immunitaires artificiels.

1.5 Conclusion

Un SMA comporte un ensemble d'agents autonomes qui interagissent les uns avec les autres à travers un environnement partagé. Son principal avantage est sa propriété qui consiste à générer des comportements émergents, en partant des interactions individuelles (par des phénomènes tels que l'auto-organisation).

Les algorithmes d'apprentissage automatique sont largement adaptés dans le domaine médical pour leur capacité d'apprentissage rapide. Ils ont montré des améliorations révolutionnaires dans plusieurs domaines des soins médicaux, en particulier dans l'analyse des données médicales. Cependant, ces méthodes souffrent encore de certains défis, d'une médecine factuelle en constante évolution, de données mitigées et de valeurs aberrantes trompeuses. Pour y parvenir, les méthodes d'optimisations sont un outil fort pour la manipulation des données, l'extraction des informations pertinentes et la résolution des problèmes difficiles.

Dans ce chapitre, nous avons présenté les principaux concepts de notre travail : SMA, apprentissage automatique et métaheuristiques. Le prochain chapitre sera dédié à la présentation de différents travaux proposés dans la littérature appliquant les techniques de l'IA dans le secteur médical.

Chapitre 2

L'IA pour la prédiction médicale : Revue de la littérature

2.1 Introduction

La médecine traverse une révolution qui cherche à répandre la virtualisation dans toutes les pratiques médicales. Cette révolution émerge de la convergence de la biologie, la médecine, les mathématiques et l'informatique avec sa capacité d'analyser de très grandes masses de données dites « big data », de déployer ces informations dans les réseaux commerciaux et sociaux et de créer des appareils numériques grand public mesurant les informations personnelles, d'où la naissance du concept médecine 4P [38].

Dans ce chapitre, nous allons définir les concepts de la médecine 4P et le PHM, décrire les travaux réalisés dans la littérature, faire une synthèse récapitulative, discuter les enjeux de l'application de l'IA pour la prédiction médicale et enfin, finir par une conclusion.

2.2 Médecine 4P

Avec l'évolution de l'outil informatique, l'augmentation exponentielle des masses de données et la multitude des types et des sources de ces données, la médecine 4P ne cesse d'apporter de nouvelles perspectives prometteuses pour le développement du secteur médical. Elle se situe à l'intersection des systèmes de la biologie et de la médecine, la révolution numérique et l'association

des patients aux réseaux sociaux et commerciaux (figure 2.1 [39]).

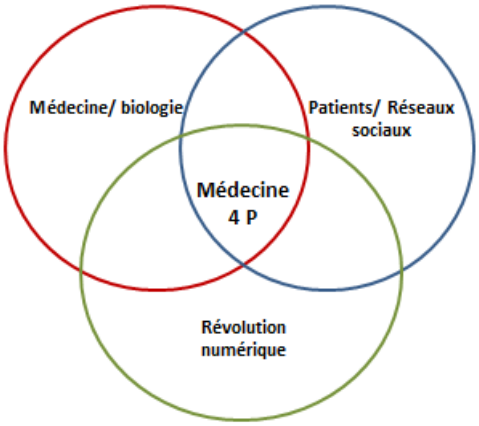


FIGURE 2.1. Contexte de la médecine 4P

Les systèmes biologiques s'appuient sur de multiples paramètres de l'individu incluant : les données génétiques, moléculaires, génomiques, les bio-marqueurs, etc. pour l'analyse et la prédiction de leur comportement (figure 2.2 [39]). Les systèmes médicaux utilisent des méthodes complexes telles que le séquençage d'ADN et d'ARN pour générer d'énormes quantités de données qui servent à la compréhension de la biologie de l'individu [40].

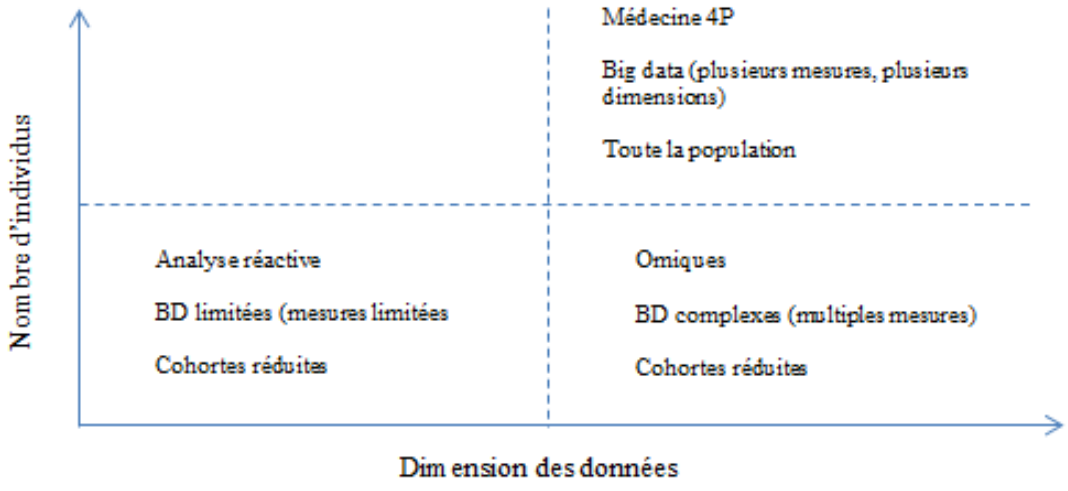


FIGURE 2.2. Evolution des données

Pour implémenter un système de médecine 4P, les données sont extraites des bases de données médicales qui peuvent contenir, en plus des molécules biologiques telles que les ADN, les ARN, les

protéines et les métabolites, les données cliniques et les données sur le mode de vie. Cela intégrera l'autosurveillance et la participation des patients à la prise de décision clinique [38]. Les différentes sources de données sont illustrées dans la figure 2.3.

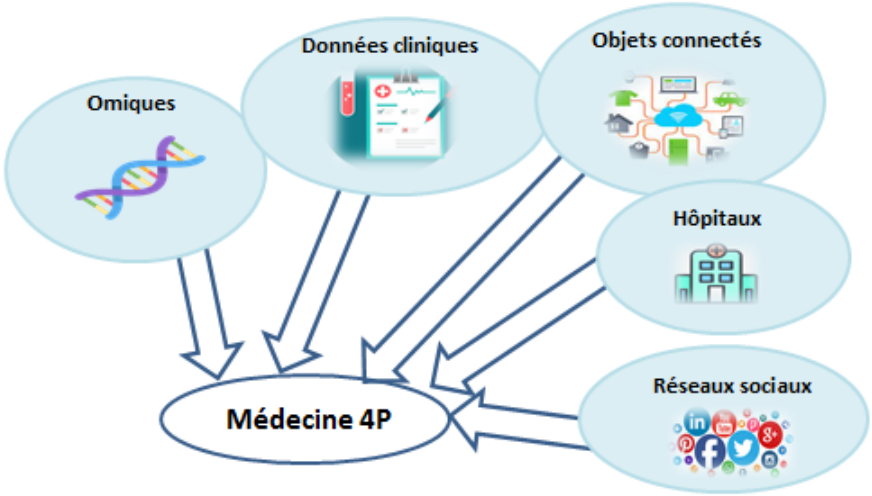


FIGURE 2.3. Les différentes sources de données

La médecine 4P est une médecine personnalisée, préventive, prédictive, et participative. Elle emploie les stratégies des systèmes de médecine et de biologie pour réaliser ces principaux objectifs :

1. Bien cibler les maladies et pronostiquer la réponse d'un patient à son traitement.
2. Assurer le bien-être des individus.
3. Progresser dans la détection des pathologies, et éviter de soumettre les malades à des examens intrusifs.
4. Faire un bon diagnostic et proposer des stratégies thérapeutiques plus adaptées aux besoins du patient, son environnement et son mode de vie.

2.2.1 La médecine prédictive

La médecine prédictive est un domaine de la médecine qui estime la probabilité de survenue d'une maladie dans le futur en tenant compte des facteurs de risque pertinents tels que l'âge, le sexe et les données cliniques mesurées, etc. Elle s'adresse aux individus ayant une prédisposition

biologique à certaines maladies en prévoyant leur apparition bien avant l'expression des symptômes [41].

Une consultation prédictive nécessite un processus multi-étapes, une interaction entre des professionnels de différentes disciplines et un temps de réflexion important[42]. L'objectif de la médecine prédictive est de prévenir l'émergence de la maladie chez les individus sains en détectant, dès le début ou avant la naissance la susceptibilité à la maladie qui peut se développer plus tard grâce au développement de biomarqueurs prédictifs [43].

La médecine prédictive est confrontée à deux défis majeurs : la communauté scientifique et médicale est loin d'avoir achevé l'inventaire fonctionnel des gènes d'une part, et d'avoir décrypté la physiopathologie de toutes les affections humaines, d'autre part. Des critiques ont émergé, faisant valoir qu'il y a un enthousiasme disproportionné pour la génétique compte tenu de ses applications actuelles dans la prédiction des risques et des traitements.

2.2.2 La médecine préventive

La médecine préventive consiste à prévenir l'occurrence d'une maladie et éviter les complications qui en résultent après son apparition chez un individu sain et sans aucun symptôme. Il existe trois niveaux de prévention et cela dépend de la l'état de santé ou du stade de la maladie du patient [44] :

1. **Prévention primaire :** qui sert à empêcher l'apparition d'une maladie en éliminant ses causes ou en augmentant le taux de résistance à cette maladie avant l'apparition des symptômes comme la vaccination contre des maladies infectieuses.
2. **Prévention secondaire :** qui sert à empêcher la propagation de la maladie aux personnes non-affectées.
3. **La prévention tertiaire :** qui sert à identifier les patients à des stades avancés et limiter cliniquement les conséquences graves par la thérapie et la réadaptation.

2.2.3 La médecine personnalisée

La médecine personnalisée utilise les informations cliniques, génétiques, génomiques et environnementales propres à chaque patient et ses maladies pour améliorer les stratégies de sa prise

en charge préventifs [45]. Elle a pour objectifs d'assurer le bon diagnostic au bon patient et au bon moment, d'anticiper sa réponse au traitement qui lui est proposé, de comprendre des maladies existantes et de développer de futurs traitements pour assurer une prise en charge individualisée.
Ce terme est né en oncologie à la fin des années quatre-vingt-dix. Elle consiste à détecter en cas d'anomalies génétiques les modifications engendrées par les signaux anormaux et déterminer les traitements appropriés [46].
La médecine personnalisée s'applique aux quatre grandes étapes du parcours de soins : la prévention, le diagnostic, le traitement et le suivi grâce au génotype de la tumeur et le génotype du patient cela permet d'adapter les doses et le traitement en fonction du métabolisme du patient [46]. La capacité de l'outil informatique joue un rôle primordial dans l'analyse du profil et des mécanismes biologiques des individus, pour une prise en charge thérapeutique et préventive efficace et adaptée aux singularités individuelles. La figure 2.4 illustre les étapes de la médecine personnalisée.

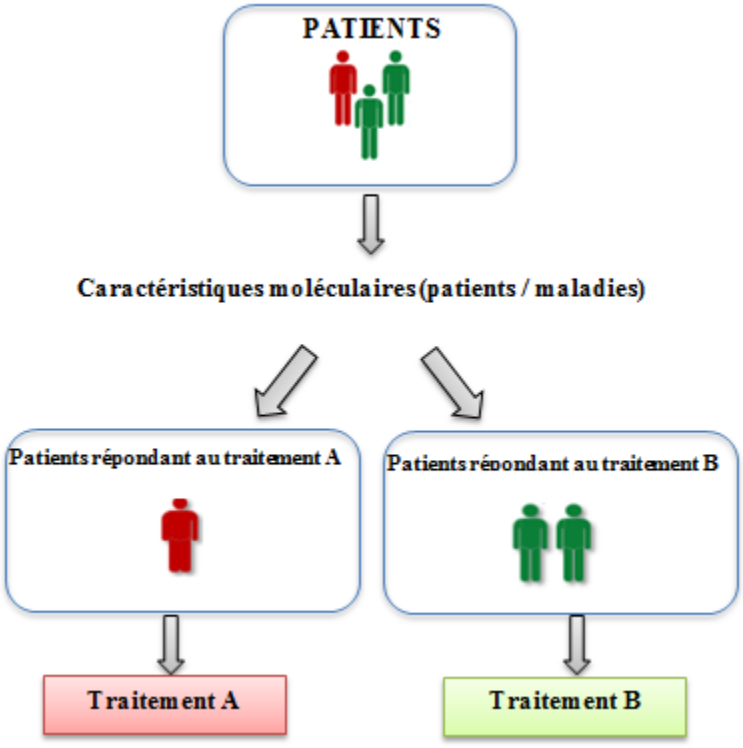


FIGURE 2.4. Médecine personnalisée

2.2.4 La médecine participative

Dans la médecine participative, le patient est acteur de son état de santé et de ses soins. Il contribue, grâce aux nouvelles technologies, à la prise en charge de sa maladie, l'évaluation de la qualité des soins cliniques et à la participation aux situations d'urgence [47].
Les réseaux sociaux et les technologies de l'information et de la communication ont un rôle important en médecine participative. Ils permettent aux individus d'entrer en contact, de consulter, de créer et d'échanger des informations en ligne sans se soucier des contraintes spatio-temporelles [48]. Le processus de la médecine participative est présenté dans la figure 2.5 .

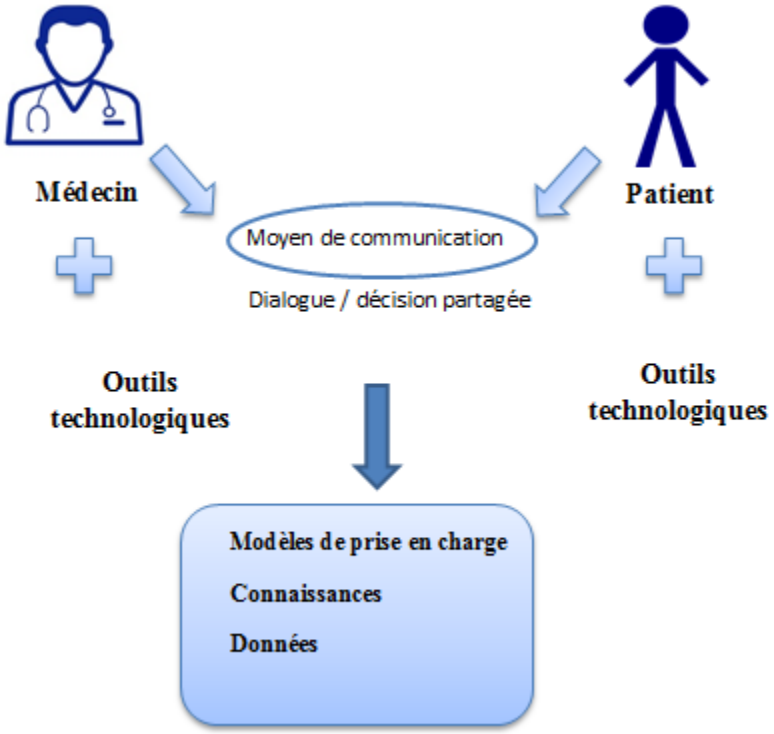


FIGURE 2.5. Médecine participative

2.3 Prognostics and health management (PHM)

Le PHM est une discipline qui permet de visualiser l'état de santé général des machines ou des systèmes complexes pour la prise de décisions concernant leur maintenance. La construction d'un PHM robuste a trois principaux objectifs :

1. L'estimation de l'état de santé actuel d'un système.

2. La prédiction d'un état futur ainsi que le temps de défaillance.
3. La détermination de l'impact d'une défaillance sur les performances d'un système.

Le PHM repose sur sept piliers : l'acquisition de données, le prétraitement des données, la détection, le diagnostic, le pronostic, la prise de décision et l'interface homme-machine. Les travaux pilotés dans la recherche PHM se concentrent sur le développement de modèles précis et robustes pour évaluer l'état de santé des systèmes en effectuant des diagnostics, des pronostics et en aidant à la prise de décision.

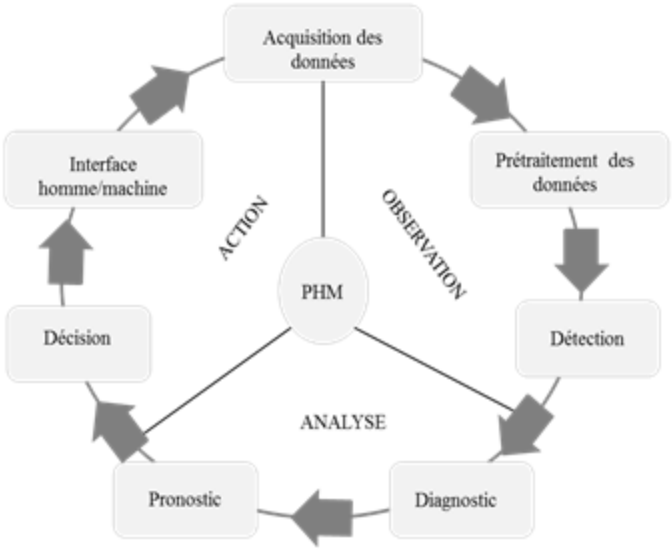


FIGURE 2.6. Processus du PHM

A partir de la figure 2.6 [49], nous pouvons remarquer que le processus du PHM est identique à celui de la médecine prédictive. En conséquence, nous pouvons l'appliquer en médecine pour détecter une éventuelle défaillance d'un organe, effectuer un diagnostic ou un pronostic d'une maladie. Néanmoins, il faut prendre en compte la complexité et la délicatesse de travailler avec un corps humain et le respect de la vie privée du patient en utilisant ses données médicales.

2.4 Revue de la littérature

Dans cette section, nous allons discuter quelques travaux récents qui traitent la thématique de la prédiction médicale sur plusieurs pathologies et en utilisant différentes méthodes.

2.4.1 Cancer du sein

Arpit B. et Aruna T. [50] ont proposé un réseau de neurones génétiquement optimisé (GONN) pour la classification des tumeurs mammaires (maligne / bénigne). Leur architecture est optimisée en introduisant un nouvel opérateur de croisement et de mutation. Pour évaluer leur approche, la base de données WOBC a été utilisée et les mesures de performance (exactitude, sensibilité, matrice de confusion, spécificité, courbe ROC et la zone sous la courbe ROC) ont été comparées avec celles du modèle classique GONN et du modèle de rétropropagation. Cette approche présente de bons résultats en termes de mesures de performance cependant, elle peut être améliorée par le prétraitement des données pour rendre GONN plus efficace pour le diagnostic en temps réel du cancer du sein.

Ashraf O. I. et Siti M. S. [51] ont proposé une méthode pour la classification automatique du cancer du sein. Ils ont appliqué le réseau de neurone MLP basé sur l'algorithme génétique NSGA-II afin d'optimiser l'exactitude et la structure de réseau de neurones. Par rapport à d'autres travaux réalisés dans la littérature, cette approche améliore l'exactitude cependant, le MLP risque de converger vers un minimum local.

Na L. et al. [52] ont proposé un modèle intelligent pour le diagnostic du cancer du sein en utilisant les bases de données WOBC et WDBC. Ce modèle est basé sur une approche hybride de sélection des données (IGSAGAW et CSSVM) pour éliminer les données redondantes et non pertinentes. Grâce à ce processus, cette méthode a montré son efficacité en améliorant l'exactitude et en réduisant la complexité et le coût de calcul.

Nawel Z. et al. [53] ont conçu un système de détection assisté par ordinateur pour la classification des mammographies. Le système proposé est basé sur les algorithmes génétiques pour la sélection d'attributs et la méthode de classification S3VM. Cette approche permet de réduire la dimensionnalité du vecteur d'entrée et d'améliorer l'exactitude. Les expérimentations ont été validées sur la base de données de mammographies DDSM.

Abdulkader H. et al. [54] ont développé un système automatique pour la classification des tissus mammaires en se basant sur deux techniques de l'apprentissage automatique BPNN et RBFN. Les tissus mammaires ont été classés en six catégories ; carcinome, fibro-adénome, mastopathie, tissu glandulaire, conjonctif et adipeux. Les données ont été acquises à l'aide de la méthode d'impédance électrique. RBFN a donné de meilleurs résultats comparé à BPNN. Cependant, les réseaux de

neurones risquent de converger vers un minimum local.

Haifeng W. et al. [55] ont conçu un modèle pour le diagnostic du cancer du sein basé sur C-SVM, a-SVM et six fonctions noyaux. Le modèle a été évalué sur les bases de données : WOBC, WDBC et SEER. Le modèle proposé augmente l'exactitude du diagnostic par rapport à d'autres travaux basés SVM. Cependant, la méthode est coûteuse en temps de calcul et en temps d'apprentissage.

Kemal P. et al.[56] ont proposé une approche hybride basée sur la normalisation par déviation absolue moyenne (MAD), la pondération des attributs basée sur KMC et classificateur AdaBoostM1. La méthode proposée a donné de bons résultats en termes d'exactitude sur la base de données BCC mais elle est coûteuse en calcul.

Teresa A. J. et al. [57] ont proposé une méthode basée sur les réseaux de neurones convolutifs (CNN) pour la classification des images de biopsie mammaire colorées à l'hématoxyline et à l'éosine (HE). Ils ont fournis quatre classes : tissu normal, lésion bénigne, carcinome in situ et carcinome invasif. L'architecture proposée a été conçue pour intégrer des informations provenant de plusieurs échelles histologiques. Le modèle a été appliqué sur une base de données issue du défi de classification de l'histologie du sein Bioimaging 2015 et composée d'images de coloration HE haute résolution non compressées et annotées.

Fabio A. et al. [58] ont utilisé une approche d'apprentissage en profondeur (DL) pour classer les images histopathologiques du cancer du sein de la base de données publique BreaKHis. Ils ont proposé une méthode basée sur l'extraction de patchs d'images pour l'entraînement du CNN et la combinaison de ces patchs pour la classification. Cette méthode permet d'éviter les adaptations du modèle qui peuvent conduire à une architecture plus complexe et coûteuse en calcul cependant, le manque de données fournit des expérimentations coûteuses.

Abdel-Zaher A.M. and Eldeib A.M. [59] ont proposé un système d'aide à la décision basé sur DBN et l'algorithme de rétro-propagation pour la détection du cancer du sein. Le système a été évalué sur la base de données WOBC et a montré une bonne performance en termes d'exactitude (99.68%) cependant, l'algorithme de rétro-propagation peut converger vers un minimum local.

Thein H.T.T and Tun K.M.M. [60] ont utilisé l'algorithme à évolution différentielle (DE) pour éviter que le réseau de neurones artificiel (ANN) converge vers un optimum local et ont présenté une approche basée sur la méthode island-based training pour la classification des tumeurs mam-

maires. L'approche a été testée sur les bases de données WDBC et WBCP et a donné de bons résultats en termes d'exactitude et de temps d'apprentissage en utilisant deux topologies de migration différentes.

Guan J.S. et al. [61] ont proposé une méthode de diagnostic basée sur le réseau de neurones SVCMAC pour la classification des tumeurs mammaires en utilisant la base de données WOBC. Les avantages de cette méthode sont : la simplicité, la rapidité de l'apprentissage et la bonne capacité de généralisation.

Kumar U.K. et al. [62] ont proposé un classificateur basé sur l'apprentissage de vote qui est une méthode ensembliste. Il combine J48, SVM et la classification naïve bayésienne (NB) sur la base de données WOBC dans le but d'améliorer les modèles de prédiction pour le système de prise de décision dans la prédiction de la capacité de survie des patients atteints de cancer du sein.

Mittal D. et al. [63] ont présenté un système de classification hybride basé sur un réseau de neurone artificiel non supervisé nommé les cartes autoadaptatives (SOM) et le classificateur supervisé de descente de gradient stochastique (SGD) pour le diagnostic du cancer du sein. Le système a été testé sur la base de données WOBC et les résultats expérimentaux ont montré qu'il y a une nette amélioration de l'exactitude par rapport aux autres travaux basés sur l'apprentissage artificiel.

Emina A. et al. [64] ont proposé un système de classification de tumeurs mammaires basé sur la régression logistique (LR),le réseau de neurone RBFN, SVM, arbres de décision, réseau bayésien, RF, rotation forest et l'algorithme génétique (GA) pour la sélection des caractéristiques . Le système a été évalué en utilisant les bases de données WBC et WOBC et la combinaison rotation forest avec GA basé 14 attributs a donné la meilleure exactitude par rapport aux autres combinaisons.

Zheng B. et al. [65] ont développé un système hybride basé sur les algorithmes k-moyennes et k-svm pour la classification des tumeurs mammaires. La méthode a été testée sur la base de données WDBC et a donné une exactitude de 97.38% en utilisant la méthode de validation croisée à 10 blocs.

Janghel R.R. et al. [66] ont proposé un système de diagnostic, pronostic et prédiction du cancer du sein pour assister les médecins. Ils se sont basés sur quatre modèles de réseaux de neurones pour implémenter leur système : MLP, RBF, LVQ et l'apprentissage compétitif (CL). L'étude expérimentale a montré que LVQ est meilleure par rapport aux autres modèles. Cependant, le modèle a été testé sur une seule base de données avec un nombre d'attributs limité.

Chaurasia V. et Pal S. [67] ont développé un système de prédiction du taux de survie des patients atteints de cancer du sein en appliquant et en comparant trois techniques de data mining : RepTree, RBF et régression logistique. Les données utilisées dans cette étude comportent 10 attributs et un total de 286 instances fournies par le Centre médical universitaire, Institut d'oncologie, Ljubljana, Yougoslavie.

Nilashi M. et al. [68] ont développé un système à base de connaissances pour la classification des tumeurs mammaires en utilisant l'algorithme espérance-maximisation (EM), les arbres de classification et de régression (CART) et l'analyse en composantes principales (PCA). Le système proposé peut être utilisé comme système d'aide à la décision clinique pour assister les médecins.

Nguyen C. et al. [69] ont proposé un système de diagnostic assisté par ordinateur (CAD) pour classer les tumeurs mammaires (bénigne / maligne). Le fonctionnement du système est assuré en se basant sur deux étapes : (a) prétraitement de la base d'entrée en éliminant les caractéristiques et attributs non pertinents et (b) la classification en utilisant l'algorithme d'apprentissage supervisé RF. Les auteurs ont testé leur système sur deux bases de données WDBC et WBCP. Cette méthode peut être appliquée à d'autres pathologies et sur d'autres bases de données. Cependant, RF devient lent et inefficace pour les prédictions en temps réel lorsqu'un grand nombre d'arbres sont générés.

Dheeba J. et al. [70] ont proposé une nouvelle approche de classification pour la détection des anomalies mammaires dans les mammographies numériques en appliquant le réseau de neurones PSOWNN. Le principe de cette approche est basé sur l'extraction des mesures d'énergie de texture des lois des mammographies et la classification des régions suspectes en utilisant un pattern classifier et en l'appliquant sur une base de données cliniques réelle (216 mammographies). L'inconvénient de cette approche est la difficulté de trouver les paramètres de conception optimaux.

Wang D. et al. [71] ont proposé une approche basée sur l'apprentissage en profondeur (DL) pour la détection des tumeurs mammaires métastatiques à partir d'images de diapositives entières de ganglions lymphatiques sentinelles. Cette approche a été testée sur la base de données Camelyon16. Ce travail a amélioré la reproductibilité, l'exactitude et la valeur clinique des diagnostics pathologiques. Cependant, il est coûteux en temps de calcul.

Wang S. et al. [72] ont développé une méthode d'extraction de règles améliorée IRFRE basée sur la forêt aléatoire (RF) pour dériver des règles de classification à partir d'un ensemble d'arbres de décision pour le diagnostic du cancer du sein. La méthode proposée a été évaluée sur les bases

de données WDBC, WOBC et SEER.

Mesut T. et al. [73] ont combiné des modèles de DL (CNNs et modèle de réseau d'encodeur automatique) pour classer les carcinomes canalaires invasifs. Les CNN AlexNet, GoogLeNet, ResNet-50, VGG-19 ont été utilisés pour l'extraction des caractéristiques pertinentes et l'apprentissage par transfert a été appliqué pour l'entrainement des CNN. Pour évaluer leur proposition, les auteurs ont utilisé une base de données publique de l'Université Case Western Reserve (277 524 patchs de 162 patients).

Mesut T. et al. [74] ont développé BreastNet, un modèle de classification basé sur un CNN. Ce modèle comprend le module attention, la technique de l'hypercolonne, le bloc résiduel. La classification a été réalisée en utilisant des images histopathologiques de tumeurs mammaires. Le modèle BreastNet a atteint une exactitude de 98,80% sur les données BreakHis composée de 7909 images histologiques.

Abdar M. et Makarenkov V. [75] ont proposé une méthode de data mining pour la prédiction du cancer du sein en appliquant les méthodes SVM et ANN pour analyser la base de données WBCD. Le modèle comprend deux techniques d'apprentissage de vote ensembliste qui aident à améliorer les performances du modèle final. De plus, les attributs pertinents de la BD et les valeurs optimales de certains paramètres importants de SVM sont identifiés pour surmonter le problème de surapprentissage.

Abir A. et Tchier F. [76] ont conçu un système de diagnostic assisté par ordinateur en combinant l'algorithme génétique évolutionnaire (EGA) et la logique floue sur la base de données saoudienne de diagnostic du cancer du sein (elle comporte les données de 260 patients) et similaire à WBCD. Le système est utilisé pour aider les médecins dans la détection précoce du cancer du sein.

Reza R. et al. [77] ont développé un système de diagnostic assisté par ordinateur pour DCE-IRM du sein sur une base de données réelle de 112 patients. Le système est basé sur un ensemble mixte de réseaux de neurones convolutifs ME-CNN pour classer les tumeurs mammaires bénignes et malignes. Le processus de diagnostic consiste en : i) segmentation tumorale basée sur l'intensité des masses et des informations morphologiques. ii) classification des tumeurs par les CNN.

Budak Ü. et al. [78] ont proposé un système de diagnostic du cancer du sein de bout en bout basé sur le réseau de neurones entièrement convolutif (FCN) pour l'extraction de caractéristiques et réseau de neurones récurrent bidirectionnel à longue mémoire à court terme (Bi-LSTM) pour la

détection des tumeurs mammaires. Les expériences ont été menées sur la base de données publique BreaKHis.

Ekici S. et Jawzal H. [79] ont développé un logiciel pour le dépistage automatique précoce du cancer du sein. Le processus est basé sur des techniques et des algorithmes pour analyser les images thermiques. La classification des images est assurée à l'aide de réseaux de neurones convolutifs (CNN) optimisés par l'algorithme de Bayes. Un taux d'exactitude de 98,95 % a été obtenu pour la base d'images thermiques appartenant à 140 individus.

Jafari-Marandi R. et al. [80] ont présenté un réseau de neurones artificiels LSSOED pour le diagnostic du cancer du sein. La méthode combine l'apprentissage supervisé et non supervisé de l'ANN sur les bases de données WOBC et WDBC. L'approche a amélioré la qualité de la prise de décision en minimisant les coûts de mauvaise classification.

Liu S. et al. [81] ont utilisé l'algorithme d'apprentissage K2 et des méthodes de calcul statistique pour construire une approche de modélisation de réseau bayésien (BN) pour la classification des tumeurs mammaires et l'aide à la prise de décision. Les données utilisées ont été collectées à partir d'une base de données d'échographies cliniques issue d'un hôpital chinois local et d'une base de données de cytologie par aspiration à l'aiguille fine (FNAC).

Liu N. et al. [82] ont proposé une approche de diagnostic du cancer du sein basée sur l'algorithme IGSAGAW pour la sélection de caractéristiques. La méthode proposée a réduit la complexité de l'algorithme SAGASW grâce à l'extraction du sous-ensemble de caractéristiques optimal. Ce qui a permis d'optimiser l'exactitude de classification maximale et le coût de mauvaise classification minimal.

Sahu B. et al. [83] ont combiné la méthode PCA et ANN pour classer les tumeurs mammaires. La méthode hybride a été appliquée sur la base de données WBCD. Comparée à d'autres algorithmes de classification, la méthode proposée a donné une bonne performance en termes d'exactitude, de sensibilité et de mesure F1.

Dans [84], L'objectif de Jabbar M.A. est de construire un système d'aide à la décision sur la base de données WDBC en utilisant le réseau bayésien (BN) et la fonction de base radiale (RBF) pour la classification des tumeurs mammaires. Ce modèle a atteint une exactitude de 97,42%.

TABLE 2.1. Résumé des caractéristiques des différents travaux traitant le BC

Critères / Travaux	Objectif	Technique	Entrées	Mesures de performance									Remarques
				1	2	3	4	5	6	7	8	9	
Arpit B. et Aruna T. 2015 [50]	Classification des tumeurs mammaires	GONN	WOBC	✓	-	✓	✓	✓	✓	✓	-	-	* Bonne exactitude * Petite base de données
Ashraf O. I. et Siti M. S. 2018 [51]	Classification des tumeurs mammaires	MLP + NSGA II	WBCD	✓	-	✓	✓	-	-	-	-	-	* Réseau optimisé * MLP peut converger vers un minimum local
Na L. et al. 2019 [52]	Classification des tumeurs mammaires	ML + GA	WDBC WOBC	✓	-	✓	✓	✓	-	-	-	✓	* Bonne performance * Coûteux en temps de calcul
Nawel Z. et al. 2016 [53]	Classification de mammographies	GA+ SSVM	Base d'images	✓	-	-	-	-	-	-	-	-	* Bonne exactitude et temps de calcul réduit. * Peut converger vers un minimum local
Abdulkader H. et al. 2017 [54]	Classification des tumeurs mammaires	ML	données par impédance électrique	✓	-	-	-	-	-	-	-	✓	* Bonne performance * Bonne capacité de généralisation * Coûteux en temps de calcul

Haifeng W. et al. 2017 [55]	Classification des tumeurs mammaires	ML	WOBC WDBC SEER	✓	-	✓	✓	✓	✓	✓	-	-	* Bonne exactitude * Coûteux en temps de calcul
Kemal P. et al. 2018 [56]	Classification des tumeurs mammaires	ML hybride	BCC	✓	✓	✓	✓	✓	✓	✓	✓	-	* Bonne exactitude * Coûteux en temps de calcul
Teresa A.J. et al. 2017 [57]	Classification d'images	CNN	Base d'images	✓	-	✓	-	-	-	-	-	-	* Bonne exactitude * Coûteux en temps de calcul
Fabio A. et al. 2016 [58]	Détection	CNN	BreaKHis	✓	-	-	-	-	-	-	-	-	* Bonne exactitude *Coûteux en temps de calcul
Abdel-Zaher A.M. et Eldeib A.M. 2016 [59]	Diagnostic	DL	WOBC	✓	-	✓	✓	✓	-	-	-	-	* Bonne exactitude *Coûteux en temps de calcul
Thein H.T.T. et Tun K.M.M. 2015 [60]	Classification des tumeurs mammaires	ANN+ DE	WDBC, WBCP	✓	-	-	-	-	-	-	-	✓	* Bonne exactitude, efficacité et fiabilité. * Coûteux en temps de calcul
Guan J.S. et al. 2016 [61]	Diagnostic	RNA SVCMAC	WOBC	✓	-	✓	✓	-	-	-	-	-	* Apprentissage rapide, * Capacité de généralisation * Calcul simple

Kumar U.K. et al. 2017 [62]	Classification des tumeurs mammaires	ML hybride	WOBC	✓	-	✓	✓	-	✓	✓	-	-	* Bonne exactitude, * Problème de surapprentissage
Mittal D. et al. 2015 [63]	Diagnostic	SOM+ SGD	WOBC	✓	-	-	-	✓	-	-	-	-	* Bonne exactitude, * Entraînement lent
Emina A. et al. 2015[64]	Classification des tumeurs mammaires	ML + GA	WOBC WDBC	✓	-	-	-	-	✓	✓	✓	-	* Bonne exactitude *Coûteux en temps de calcul
Zheng B. et al.2013 [65]	Diagnostic	K-NN + SVM	WDBC	✓	-	-	-	-	-	✓	-	✓	* La méthode proposée n'est pas scalable (grande BD avec des valeurs manquantes est un défi)
Janghel R.R. et al. 2014 [66]	Diagnostic	RNN (4 méthodes)	WBCD	✓	-	-	-	✓	-	-	-	✓	* La méthode proposée n'est pas scalable
Chaurasia V. et Pal S. 2014 [67]	Diagnostic pronostic (prédiction de survie)	Data mining (3 méthodes)	Données collectées	✓	-	-	-	-	-	-	-	✓	* Petite base de données

Nilashi M. et al. 2017 [68]	Diagnostic	EM + logique floue + PCA + CART	WBCD	✓	-	-	-	-	✓	✓	-	-	* EM échoue sur des BD de grandes dimensions
Nguyen C. et al. 2019 [69]	Diagnostic /pronostic	RF + sélection d'attributs	WDBC WBCP	✓	-	✓	✓	✓	✓	✓	-	-	* RF devient lent et inefficace pour les prédictions en temps réel lorsqu'un grand nombre d'arbres sont générés
Dheeba J. et al. 2014 [70]	Diagnostic	PSOWNN	Base d'images réelles	✓	-	✓	✓	-	✓	✓	-	-	* Bonne performance * Difficulté de trouver les paramètres de conception optimaux.
Wang D. et al. 2016 [71]	Détection des tumeurs mammaires métastatiques	DL	Camelyon16 dataset	✓	-	✓	-	-	✓	✓	-	-	* Coûteux en temps de calcul
Wang S. et al. 2019 [72]	Diagnostic	RF	WDBC WOBC SEER	✓	-	✓	✓	-	-	-	-	-	* Bonne exactitude et interprétabilité * Inefficacité de la phase d'entrainement

Mesut T. et al. 2020 [73]	Diagnostic	DL	Base d'images réelles	✓	✓	✓	✓	✓	✓	✓	✓	-	* Bonne performance * Coûteux en temps de calcul
Mesut T. et al. 2020 [74]	Diagnostic	DL	BreakHis	✓	✓	✓	✓	✓	✓	✓	✓	-	* Bonne performance * Coûteux en temps de calcul
Abdar M. et Makarenkov V. 2019 [75]	Diagnostic	ML	WBCD	✓	✓	✓	✓	-	✓	✓	-	-	* Bonne exactitude * Le problème de surapprentissage est surmonté * Coûteux en temps de calcul
Abir A. et Tchier F. 2017 [76]	Diagnostic	Logique floue + EGA	WBCD BD saoudienne du cancer du sein	✓	-	-	-	-	-	-	-	-	* Bon degrés de confidence 91% * Coûteux en temps de calcul
Reza R. et al. 2017 [77]	Diagnostic	ME-CNN	Base d'images réelles	✓	-	✓	✓	-	✓	✓	-	✓	* Efficace en temps d'exécution * Petite BD
Budak Ü. et al. 2019 [78]	Diagnostic	FCN + Bi-LSTM	BreakHis	✓	-	✓	✓	-	-	-	-	-	* Bonne exactitude * Coûteux en temps de calcul

Ekici S. et Jawzal H. 2020 [79]	Dépistage automatique précoce	CNN + Algorithme de Bayes	Base d'images réelles	✓	✓	-	-	✓	-	-	-	-	* Bonne exactitude * Petite BD
Jafari-Marandi R. et al. 2018 [80]	Diagnostic	RNN supervisé et non supervisé	WOBC, WDBC	✓	✓	✓	✓	-	-	-	-	✓	* Bonne exactitude * Coûteux en temps de calcul
Liu S. et al. 2018 [81]	Diagnostic	BN + K2 learning algorithm + statistical computation methods	FNAC base de données d'échographie clinique	✓	-	✓	✓	-	✓	✓	-	-	* Peut être appliqué pour le diagnostic d'autres pathologies *Coûteux en temps de calcul
Liu N. et al. 2019 [82]	Diagnostic	IGSAGAW + CSSVM	WBCD WBC	✓	-	✓	✓	✓	-	-	-	✓	* Amélioration de la exactitude et du temps de calcul * Complexité élevée.
Sahu B. et al. 2019 [83]	Diagnostic	PCA + ANN	WBCD	✓	✓	✓	✓	-	✓	✓	✓	-	* Bonne performance * Faible interprétabilité
Jabbar M.A. 2021 [84]	Classification des tumeurs mammaires	ML techniques	WOBC	✓	✓	✓	✓	✓	✓	✓	✓	✓	* Evalué sur une seule dataset

(1 : exactitude, 2 : valeur prédictive positive, 3 : sensibilité, 4 : spécificité, 5 : matrice de confusion, 6 : la courbe ROC, 7 : la zone sous la courbe ROC, 8 : score F1, 9 : temps de calcul)

2.4.2 Maladies cardiovasculaires

Les auteur de cet article [85] ont proposé un modèle prédictif pour la détection des maladies cardiovasculaires (CVDs) en se basant sur les techniques d'apprentissage automatique (ML) et data mining. L'approche proposée combine NB et GA pour classer les CVDs. Les données ont été collectées à partir de la base de données sur les CVDs de Cleveland (CHDD).

Subanya B. et Rajalaxmi R.R. [86] ont utilisé un algorithme méta-heuristique (colonie d'abeilles) pour déterminer le sous-ensemble de caractéristiques optimales avec une meilleure exactitude de classification pour le diagnostic des CVDs. Les données ont été extraites du référentiel UCI (une base de données des CVDs).

Singh M. et al. [87] ont appliqué le modèle d'équation structurelle (SEM) pour identifier les relations entre les variables considérées comme liées à la cause des CVDs et la carte cognitive floue (FCM) pour évaluer les résultats obtenus dans un système prédictif qui aide à la détection des personnes à risque de développer des CVDs. Dans cette étude, les données ont été extraites de la source Canadian Community Health Survey (CCHS). Cependant, le nombre d'attributs utilisés n'est pas suffisant pour avoir un modèle très précis.

Narain R. et al. [88] ont proposé un système pour la prédiction des CVDs utilisant un réseau de neurones quantiques (QNN). Les données de 689 patients présentant des symptômes de CVD et une base de données de 5209 patients CVD de l'étude de Framingham ont été utilisées. Ce système a été évalué et comparé au score de risque de Framingham (FRS) et a montré de bons résultats en termes d'exactitude.

Venkatalakshmi B. et al. [89] ont développé un système de diagnostic et de prédiction de CVDs. Le fonctionnement de ce système repose sur deux algorithmes : arbre de décision (DT) et la classification naïve de Bayes (NB). La base de données utilisée pour l'évaluer se compose d'attributs et de valeurs collectés à partir du référentiel d'apprentissage automatique UCI. Afin d'améliorer l'efficacité et l'exactitude, un algorithme génétique de processus d'optimisation a été utilisé. Pour améliorer la performance de ce système, un prétraitement de la base d'entrée est requis.

TABLE 2.2. Résumé des caractéristiques des différents travaux traitant les CVDs

Critères / Travaux	Objectif	Technique	Entrées	Mesures de performance									Remarques
				1	2	3	4	5	6	7	8	9	
Makwana A. et Patel J. 2015 [85]	Detection de CVDs	ML + Data Mining	CHDD	-	-	-	-	-	-	-	-	-	* Le système n'a pas été évalué
Subanya B. et Rajalaxmi R.R. 2014 [86]	Classification des CVDs	méta-heuristique (colonie d'abeilles)	BD de CVDs	✓	-	-	-	✓	-	-	-	-	* Facile à implémenter
Singh M. et al. 2016 [87]	Modèle prédictif pour les risques des CVDs	SEM+ FCM	CCHS	✓	-	✓	✓	-	✓	✓	-	-	* En utilisant plus d'attributs l'exactitude peut être améliorée
Narain R. et al. 2016 [88]	Prédiction du risque de CVDs	QNN	* Données d es patients * BD de l'étude Framingham	✓	-	-	-	-	-	-	-	-	* Testé sur multiples BD * Coûteux en calcul
Venkatalakshmi B. et al. 2014 [89]	Prédiction de CVDs	DT + NB	Données collectées	✓	-	✓	✓	-	-	-	-	-	* Petite BD

(1 : exactitude, 2 : valeur prédictive positive, 3 : sensibilité, 4 : spécificité, 5 : matrice de confusion, 6 : la courbe ROC, 7 : la zone sous la courbe ROC, 8 : score F1, 9 : temps de calcul)

2.4.3 Covid-19

Ahmed H. et al. [90] ont proposé une nouvelle variante de l'algorithme KNN (KNNV) pour le diagnostic de la COVID-19. Les résultats expérimentaux sur la base de données COVID-19 incomplète et hétérogène (IHC) issue de la société italienne de radiologie médicale et d'intervention ont montré une bonne performance en termes de quatre métriques : exactitude, précision, rappel et score F1.

Tulin O. et al. [91] ont conçu un modèle basé sur le DL pour la détection automatique de la COVID-19 en utilisant des images radiographiques thoraciques brutes et le modèle DarkNet.

Chimmula V.K.et Zhang L [92] ont développé un réseau basé sur LSTM pour prédire la transmission de la COVID-19 en temps réel. Les données utilisées sont des séries chronologiques de l'Université Hopkins et de l'Autorité canadienne de la santé. Le modèle proposé est performant avec une exactitude de 92,67 % pour les prédictions à long terme.

Jordi L. et al. [93] ont développé un système de traitement de la parole par IA pour le diagnostic non invasif et en temps réel de la COVID-19 à partir d'un ensemble d'enregistrements de toux. Le système proposé est basé sur les CNN entrainés sur 4256 sujets et testés sur les 1064 sujets restants de la BD utilisée.

Marques G. et al. [94] ont proposé un système d'aide à la décision médicale basé sur le CNN utilisant l'architecture EfficientNet. Ils ont présenté deux expériences : classification multi-classe et classification binaire. Le système a été développé en utilisant une base d'images radiographiques.

Wang S. et al. [95] ont proposé un système basé sur le CNN pour le diagnostic de la COVID-19. Ils ont collecté des images CT de 259 patients (180 cas de pneumonie virale typique et les 97 cas de trois hôpitaux différents avec un sras-cov-2 confirmé) pour la mise en œuvre de leur système. Ils ont atteint une exactitude de 89,5%. Cependant, il existe certaines limites à leur étude : la classification des images CT est difficile et la base d'apprentissage est petite.

Babukarthik R.G. et al. [96] ont proposé une méthode de DL basée sur le réseau de neurones convolutifs d'apprentissage profond génétique (GDCNN). Le but de cette étude est de fournir une solution pour identifier les poumons sains des poumons atteints de pneumonie due à la COVID-19. Ils ont utilisé des images CXR (plus de 5000 images) pour développer leur modèle et l'ont comparé avec ReseNet18, ReseNet50, Squeezenet, DenseNet-121 et Visual Geometry Group (VGG16). La méthode proposée a prouvé sa performance, néaumoins, elle est couteuse en termes de compléxité

de calcul.

Moutaz A. et al. [97] ont proposé un système intelligent pour la détection et la prévision de la COVID-19. Leurs principaux objectifs sont : (a) distinguer les patients COVID-19 des individus sains sur la base d'images radiographiques thoraciques utilisant le réseau de neurone convolutif VGG 16 ; (b) prédire les confirmations COVID-19, les rétablissements et les décès au cours des 7 prochains jours ; et (c) trouver les zones les plus touchées en appliquant trois méthodes de prévision : ARIMA, LSTM, algorithme Prophet (PA). Ils ont utilisé une base d'images réelles (128 images radiographiques thoraciques / 1000 après augmentation des données). Les résultats ont donné des performances prometteuses (exactitude de 94,80 % et 88,43 % en Australie et en Jordanie, respectivement) et l'algorithme PA a montré ses performances par rapport à LSTM et ARIMA pour les prévisions COVID-19 pour les 7 prochains jours.

Alakus T.B. et al. [98] ont réalisé un modèle prédictif clinique qui estime la probabilité de développer la COVID-19 en utilisant l'apprentissage en profondeur (hybride CNN et LSTM) et des données de laboratoire (la BD contient 111 analyses de laboratoire provenant de 5644 patients différents). Le modèle a montré de bonnes performances en termes de exactitude de 86,66 %, de score F1 de 91,89 %, de précision (PPV) de 86,75 %, de rappel de 99,42 % et d'une zone sous la courbe ROC de 62,50 %. Cependant, la base de données utilisée était petite et déséquilibrée.

Song Y. et al. [99] ont proposé une approche de DL pour aider les médecins à détecter la COVID-19 et à identifier automatiquement les lésions à partir d'images CT. Les données ont été fournies par l'hôpital Renmin de l'université de Wuhan, et l'hôpital mémorial Sun Yat-Sen de l'université Sun Yat-sen de Guangzhou (images CT de 88 patients avec la COVID-19, 101 patients infectés par une pneumonie bactérienne, et 86 personnes en bonne santé). Pour évaluer l'efficacité de l'architecture DRENet, elle a été comparée avec Resnet, DenseNet et VGG16. L'approche proposée a donné de bons résultats en termes d'exactitude, zone sous la courbe ROC, PPV, score F1, et le rappel qui sont respectivement de 0,94, 0,99, 0,96, 0,94 et 0,93.

Salman F.M. et al. [100] ont conçu et mis en œuvre un réseau de neurones convolutif profond basé sur l'algorithme Inception V3 pour la prédiction de la COVID-19. Ils ont utilisé une base de données de 260 images radiographiques (130 covid-19, 130 normales). Pour évaluer le modèle, plusieurs mesures de performance ont été calculées (exactitude, spécificité, rappel, PPV, score F1, et le rappel). Le modèle proposé a atteint une exactitude de 100%.

Le manque de données accessibles au public est l'un des défis auxquels est confronté le diagnostic automatique de la COVID-19. Pour faire face à ce problème, Maghdid H. et al. [101] ont construit une base de données composée de (170 images radiographiques et 361 images CT). Ensuite, ils ont proposé un système de diagnostic automatic de la COVID-19 basé sur CNN et l'algorithme d'apprentissage par transfert (modèle AlexNet).

Trois algorithmes de ML ont été utilisés pour prédire le temps après lequel le nombre de nouveaux cas cesse d'augmenter. Pour cela, Khan F.M. et al. [102] se sont basés sur des données rapportées quotidiennement (cas confirmés, décédés et rétablis) en Inde. Pour évaluer la performance des trois algorithmes (SVM, arbre de décision et l'algorithme de régression par processus gaussien (GPR)), l'erreur quadratique moyenne (RMSE), l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE) et le coefficient de détermination (R2) ont été calculés.

Wu J. et al. [103] visaient à concevoir un outil de pré-alerte plus pratique que la PCR et le test sérologique pour la détection précoce du COVID-19 et le dépistage initial des patients suspects. L'outil a été construit sur la base de 11 caractéristiques (indices sanguins cliniques) qui ont été extraites à l'aide de l'algorithme de forêt aléatoire. Le système a montré des résultats prometteurs en termes d'exactitude, de sensibilité (rappel), de spécificité, de coefficient de corrélation de Matthews (MCC) et de la zone sous la courbe (AUC). Cependant, son application sur l'identification des cas de COVID-19 présentant des symptômes atypiques n'est pas encore confirmée.

Nemati M. et al.[104] ont proposé un modèle pour analyser les caractéristiques de survie de 1 182 patients et prédire le temps de sortie des patients hospitalisés. Les résultats, en utilisant la mesure C-index, indiquent que le modèle de l'amplification de gradient (Gradient Boosting) dépasse les autres modèles de prédiction de la survie des patients.

Hu S. et al. [105] ont proposé une approche d'apprentissage en profondeur faiblement supervisée pour la classification COVID-19 et la détection des lésions en utilisant 150 examens volumétriques 3D du thorax (cas de pneumonie acquise dans la communauté CAT et cas de non-pneumonie NP), ils ont également utilisé la base de données TCIA pour former le réseau. Le modèle proposé a atteint une bonne performance pour la classification et la détection des lésions.

Ardakani A.B. et al. [106]ont proposé un système de diagnostic assisté par ordinateur. Les auteurs ont utilisé 1020 images CT pour distinguer entre la covid-19 et d'autres maladies pulmonaires. Pour cette raison, 10 CNN pré-entrainés (AlexNet, VGG-16, VGG-19, SqueezeNet, GoogleNet,

MobileNet-V2, ResNet-18, ResNet-50, ResNet-101 et Xception) ont été utilisés. Les résultats expérimentaux ont montré que le DL pouvait distinguer COVID-19 d'autres maladies de pneumonie atypiques et virales avec une bonne exactitude. Les meilleurs résultats ont été atteints grâce aux réseaux ResNet-101 et Xeption.

Barstugan M. et al.[107] ont proposé un système de diagnostic précoce de la covid-19 en utilisant quatre bases de données de 150 images CT. Le processus de détection est implémenté en combinant les algorithmes de sélection de caractéristiques suivants : Grey Level Co-occurrence Matrix (GLCM), Local Directional Pattern (LDP), Grey Level Run Length Matrix (GLRLM), Grey-Level Size Zone Matrix (GLSZM), and Discrete Wavelet Transform (DWT) avec SVM pour la classification et 2-fold, 5-fold and 10-fold cross-validations pour la validation. Les métriques : sensibilité, spécificité, exactitude, PPV et le score F1 ont été utilisées pour évaluer la performance du système. La meilleure exactitude de classification (99,68%) a été obtenue avec la combinaison de validation croisée 10-fold et GLSZM. Ces résultats montrent que la méthode proposée pourrait être utilisée pour diagnostiquer la maladie COVID-19 en tant que système assistant.

Alom M.Z. et al. [108] ont appliqué les méthodes de DL pour proposer une méthode pour la détection de bout en bout de la COVID-19 et la localisation de la zone infectée. Les méthodes proposées sont évaluées sur des images radiographiques et tomodensitométriques. Ils ont employé une approche de réseau de neurones convolutifs récurrents résiduels de démarrage (IRRCNN) avec apprentissage par transfert (TL) pour la détection de la COVID-19 et le modèle de réseau NABLA-3 pour segmenter les zones infectées. Les résultats qualitatifs et quantitatifs démontrent l'efficacité du système proposé pour la détection de la COVID-19 et la localisation des zones infectées .

Zheng C. et al.[109] ont développé un système basé sur l'apprentissage en profondeur faiblement supervisé utilisant des images CT 3D pour détecter la COVID-19. L'UNet pré-entraîné a été appliqué pour la segmentation de la région pulmonaire. 499 images CT 3D ont été utilisées pour l'entrainement du réseau et 131 images CT 3D ont été utilisées pour les tests. L'algorithme proposé a atteint 0,959 pour la zone sous la courbe ROC (ROC AUC) et 0,976 pour la zone sous la courbe precision-recall (PR AUC). Les auteurs ont obtenu une sensibilité de 0,907 et une spécificité de 0,911 dans la courbe ROC. En utilisant un seuil de probabilité de 0,5 pour classer COVID-positif et COVID-négatif, l'algorithme a obtenu une exactitude de 0,901, une précision de 0,840 et une valeur prédictive négative très élevée de 0,982.

Pinter G. et al.[110] ont proposé une approche de ML hybride basée sur un système d'inférence floue basé sur un réseau adaptatif (ANFIS) et un algorithme compétitif perceptron-impérialiste multicouche (MLP-ICA) pour prédire la COVID-19. L'objectif de cette étude est de prédire les séries chronologiques d'individus infectés et le taux de mortalité en utilisant une base de données hongroise. La validation est effectuée pendant 9 jours et les mesures de performance utilisées (coefficient de détermination, MAPE et RMSE) ont montré des résultats prometteurs, ce qui confirme l'exactitude du modèle.

L'objectif de Muhammad L. Jet al. [111] est de classer les cas positifs et négatifs de COVID-19 au Mexique en appliquant des méthodes de ML supervisé (régression logistique (LR), l'arbre de décision (DT), une machine à vecteurs de support (SVM), naïf Bayes (NB) et un réseau artificiel neutre (ANN)). Ils ont utilisé une base de données étiquetées épidémiologiques. L'exactitude la plus élevée a été obtenue par arbre de décision (94,99 %) tandis que d'autres modèles développés avec régression logistique, Bayes naïfs, SVM et ANN ont respectivement 94,41%, 94,36 %, 92,40 % et 89,20 % d'exactitude.

Panwar H. et al.[112] ont proposé une méthode de dépistage rapide (nCOVnet) basée sur un réseau de neurones profonds pour détecter les patients COVID-19 à l'aide d'images radiographiques. La base de données utilisée dans ce travail est open source et se compose d'environ 192 images radiographiques de patients positifs au COVID-19 et d'un total de 337 images. La méthode proposée est évaluée avec quatre métriques : la matrice de confusion, la courbe ROC, la zone sous la courbe ROC et l'exactitude d'apprentissage. Le système détecte correctement les patients positifs au COVID-19 avec une exactitude de 97% alors que l'exactitude globale du modèle proposé est de 88%.

Vaid S. et al.[113] ont développé un modèle de DL pour prédire avec précision la maladie à partir de radiographies pulmonaires en utilisant une base de données publique. Leur modèle est basé sur l'apprentissage par transfert, il vise à détecter les anomalies structurelles et la catégorisation des maladies. L'approche proposée a montré de bonnes performances en termes d'exactitude (96,3%). Cependant, la base de données utilisé est petite (181 cas COVID-19).

Shah S.H.H. et al.[114] ont proposé une approche basée agent pour fournir un service d'assistance mobile et de surveillance du port des masques faciaux dans les grandes surfaces de manière efficace et rentable. Le système multi-agent proposé repose sur la collaboration de deux agents :

1) Caméra CCTV et 2) Robot. Les résultats expérimentaux montrent son efficacité dans la navigation et le contrôle du robot sur la base d'informations visuelles interprétées par une caméra de vidéosurveillance en utilisant des algorithmes de vision par ordinateur en temps réel.

Dans les articles [115] et [116], les auteurs ont proposé deux approches basées sur le SMA pour simuler les maladies contagieuses. Par rapport aux approches de la littérature, l'approche proposée dans [115] ajoute un nouvel ensemble de fonctionnalités pour contrôler l'épidémie pendant la simulation afin de vérifier dynamiquement comment les stratégies gouvernementales peuvent avoir un impact sur la propagation de la maladie. La simulation est basée sur de multiples facteurs tels que la réponse des administrations locales, la présence d'entreprises locales qui ont un rôle important dans la propagation de la maladie et ses effets sur la population. Le système de simulation est utilisé pour la pandémie de COVID-19 et il présente une forte ressemblance avec les scénarios les plus réalistes. Le modèle développé dans [116] a permis de prédire la propagation de la COVID-19 dans différents pays et les résultats obtenus ont montré son efficacité .

Dans [117], les auteurs ont proposé un modèle automatisé basé sur l'unité récurrente fermée (GRU) empilée pour la prédiction de la Covid-19 en s'appuyant sur les habitudes des patients et leurs dossiers médicaux. Ils ont utilisé la base de données de patients pré-conditionnels de Kaggle et ont obtenu une exactitude de 65,36%.

Les auteurs de [118] ont conçu un système expert basé sur la logique floue (FL) pour la prédiction de l'admission en soins intensifs chez les patients COVID-19 à l'aide d'un ensemble de données covid-19 accessible au public. Ils ont comparé leur approche à NB, CBR, DT et KNN, les résultats obtenus démontrent l'efficacité de l'approche proposée avec une exactitude de 91,6% et un scor F1 de 95,%.

2.4.4 Autres pathologies

Boden L.M. et al. [119] ont proposé une méthode mathématique pour prédire la probabilité d'une intervention chirurgicale en utilisant la base d'échantillons de 8006 patients lombalgiques. Des facteurs de risque indépendants pour subir une chirurgie de la colonne vertébrale ont été identifiés grâce à une analyse statistique univariée et multivariée, et le modèle de probabilité de chirurgie de la colonne vertébrale (SSL) a été créé en utilisant un échantillon aléatoire de 80% du total des patients de la cohorte utilisée, et validé sur les autres 20%.

TABLE 2.3. Résumé des caractéristiques des différents travaux traitant la COVID-19

Critères / Travaux	Objectif	Technique	Entrées	Mesures de performance									Remarques
				1	2	3	4	5	6	7	8	9	
Ahmed H. et al. 2021 [90]	Diagnostic	KNNV	La BD IHC	✓	✓	-	-	-	-	-	✓	-	* Le modèle proposé peut être appliqué sur d'autres pathologies * Données incomplètes
Ozturk T. et al. 2020 [91]	Détection automatique	DL	Images radiographiques	✓	✓	✓	✓	✓	-	-	✓	-	Nombre d'images limité
Chimmula V.K. et Zhang L. 2020 [92]	Prédiction de la transmission	LSTM	BD de séries temporelles	✓	-	-	-	-	-	-	-	-	* Bonne exactitude * Petite BD
Jordi L. et al. 2020 [93]	Détection automatique	CNN	BD d'enregistrements vocaux	✓	-	✓	✓	-	✓	✓	-	-	* Non invasif, gratuit et en temps réel * Problème de la qualité des données d'entrée.
Marques G. et al. 2020 [94]	Diagnostic	CNN	images radiographiques	✓	✓	✓	✓	✓	✓	✓	✓	-	* La performance diminue à un stade précoce de la maladie
Wang S. et al. 2021 [95]	Diagnostic	DL	Images CT	✓	✓	✓	✓	-	✓	✓	✓	-	* Petite base d'apprentissage * Classification difficile.

Babukarthik R.G. et al. 2021 [96]	Diagnostic	GDCNN	Images CXR	✓	✓	✓	✓	✓	-	-	✓	-	* Bonne performance * Coûteux en calcul
Alazab M. et al. 2020 [97]	* Diagnostic * Prédiction de la transmission	DL	Images radiographiques	✓	✓	✓	✓	-	-	-	✓	-	* Réduit le prétraitement manuelle des données. * Petite BD
Alakus T.B. et al. 2020 [98]	Diagnostic	CNN + LSTM	BD d'analyses	✓	✓	✓	✓	-	✓	✓	✓	-	* Peut être utile pour l'aide à la décision * BD : petite et déséquilibrée
Song Y. et al. 2021 [99]	Diagnostic	CNN	Images radiographiques	✓	✓	✓	✓	✓	✓	✓	✓	-	* Bonne performance * Coûteux en calcul
Salman F.M. et al. 2020 [100]	Prédiction	DL	Images radiographiques	✓	✓	✓	✓	-	-	-	✓	-	* Très bonne performance * Le nombre d'images utilisées est limité
Maghdid H. et al. 2021 [101]	* Création d'une BD * Diagnostic	* DL * Apprentissage par transfert	BD d'images (radiographiques et CT)	✓	-	✓	✓	✓	✓	✓	-	-	* Bonne performance * Petite BD
Khan F.M. et al. 2021 [102]	Prédiction de la transmission	* ML * GIS	Données collectées	✓	-	-	-	-	-	-	-	-	Coûteux en calcul
Wu J. et al. 2020 [103]	Diagnostic précoce	ML	Base d'indices sanguins cliniques	✓	-	✓	✓	-	✓	✓	-	-	* Bonne exactitude * Petite BD * Son application n'est pas confirmée sur tous les cas.

Nemati M. et al. 2020 [104]	Prédiction de la survie	ML	Données cliniques	✓	-	-	-	-	-	-	-	-	* Bonne exactitude * Petite BD
Hu S. et al. 2020 [105]	Classification et détection des lésions	DL	Images CT	✓	✓	✓	✓	-	✓	✓	-	-	* Minimise le besoin d'étiquetage manuel des images * Le réseau est formé sur des images individuelles * Pas assez discriminant.
Ardakani A.B. et al. 2020 [106]	Diagnostic	DL	Images CT	✓	-	✓	✓	✓	✓	✓	-	-	* Petite BD * Coûteux en calcul
Barstugan M. et al., 2020 [107]	Diagnostic précoce	ML	Images CT	✓	✓	✓	✓	-	-	-	✓	-	* Nombre d'images limité * Le système doit être testé sur d'autre BD
Alom M.Z. et al., 2020 [108]	* Détection * localisation des régions infectées	DL	Images CT et radiographiques	✓	-	-	-	-	✓	✓	-	-	Petite base d'apprentissage.

Zheng C. et al. 2020 [109]	Détection	DCNN	3D CT images	✓	✓	✓	✓	-	✓	✓	-	-	Bonnes performances sans qu'il soit nécessaire d'annoter les lésions COVID-19 dans les volumes CT pour l'apprentissage
Pinter G. et al., 2020 [110]	Prédiction	ANFIS et MLP-ICA (approche hybride ML)	Rapports statistiques sur les cas de COVID-19 et le taux de mortalité en Hongrie (disponibles en ligne)	✓	-	-	-	-	-	-	-	-	* Résultats prometteurs * Problèmes de BD (incomplète, non testée)
Muhammad L.J et al., 2021 [111]	Classification des cas Covid-19	ML (méthodes d'apprentissage supervisé)	BD épidémiologique étiqueté	✓	-	✓	✓	-	-	-	-	-	* Permet de réduire les interactions patients/médecins
Panwar H. et al., 2020 [112]	Dépistage de la COVID-19	CNN (VGG 16)	Images radiographiques	✓	-	-	-	✓	✓	✓	-	✓	* Dépistage rapide (moins de 3 secondes) * La exactitude peut être améliorée avec l'augmentation des données
Vaid S. et al., 2020 [113]	Prédiction	CNN	Images radiographiques	✓	✓	✓	-	✓	-	-	✓	-	Le nombre d'images utilisées est limité (181 images)

Shah S.H.H. et al., 2021 [114]	Service d'assistance mobile	SMA	vidéos de la caméra et du robot	✓	✓	✓	-	-	-	-	-	-	Testé et validé en situation réelle
Nanna G.A. et al. 2020 [115]	Simulation de la propagation de la COVID-19	SMA	Données réelles	-	-	-	-	-	-	-	-	-	Certains aspects doivent être améliorés pour que la simulation soit plus réaliste.
Vyklyuk Y. et al., 2021 [116]	Simulation de la propagation de la COVID-19	SMA	Données réelles	-	-	-	-	-	-	-	-	-	Incapacité à comparer avec précision les résultats de la simulation avec les données réelles
Bandyopadhyay S. K. et Dutta S. 2020 [117]	Prédiction	DL (modèle Stacked-Bi-GRU)	Pre-condition patient dataset de Kaggle	✓	✓	✓	✓	-	-	-	✓	-	Coûteux en calcul
Asl A.A.S. et al., 2021 [118]	Prédiction	La logique floue	Pre-condition patient dataset de Kaggle	✓	-	-	-	-	-	-	✓	-	Ce modèle fournit de meilleurs résultats lorsque les variables d'entrée sont continues.

(1 : exactitude, 2 : valeur prédictive positive, 3 : sensibilité, 4 : spécificité, 5 : matrice de confusion, 6 : la courbe ROC, 7 : la zone sous la courbe ROC, 8 : score F1, 9 : temps de calcul)

Søreide K. et al. [120] ont proposé une approche basée sur le réseau de neurones artificiels (ANN), le perceptron multicouche (MLP) pour prédire la mortalité des patients atteints d'ulcère gastroduodénal perforé. L'entrée de ce modèle est un échantillon de patients analysés par Statistical Package for Social Sciences (IBM SPSS v. 21, Inc. pour Mac). Son principe est de proposer trois modèles de MLP et de donner le modèle optimal. Cependant, dans ce genre d'approches, l'intervention de l'expert humain est indispensable pour la collecte des données et des problèmes garbage-in, garbage out peuvent exister.

Hunt N. et al. [121] ont proposé un modèle basé sur une analyse spatiale pour la prédiction de la rage au Tennessee . La méthode proposée consiste en trois étapes :

1. Acquisition des données du ministère de la Santé du Tennessee ;
2. Traitement des données avec le logiciel ArcGIS pour obtenir le modèle prédictif ;
3. Analyse spatiale avec les logiciels Fragstats et Circuitscape.

Le système a réalisé un ensemble de modèles (cartes) tels que des modèles de distribution, un modèle de densité, etc. Cependant, il ne permet pas de surveiller la maladie en temps réel et il devient inefficace lorsque le volume de l'ensemble de données augmente.

Devi C.S. et al.[122] ont décrit un système distribué d'e-santé pour le diagnostic automatique de l'état d'un patient à partir de ses données sans l'intervention d'un médecin. Ce service est disponible sur Internet. Lorsque l'état d'un patient change, le système alerte automatiquement le médecin. Cela est mis en œuvre à l'aide d'un système multi-agents (MAS) et d'un système à inférence neuro-flou adaptatif (ANFIS). Les différents agents du système sont répartis sur différents sites et communiquent d'une façon asynchrone pour atteindre leur objectif global.

Das K. et al. [123] ont présenté une approche qui combine l'algorithme génétique, les algorithmes de recherche d'harmonie (HAS) et SVM pour la sélection de gènes informatifs. Cependant, les méthodes heuristiques dépendent du problème et elles tendent généralement vers un optimum local qui ne permet pas d'obtenir la solution globale optimale. Le modèle proposé a été appliqué sur plusieurs BD de différentes pathologies et a été évalué en utilisant des mesures probabilistes.

Sahebi G. et al. [124] ont proposé une méthode de sélection de caractéristiques basée sur l'algorithme génétique. Pour évaluer les sous-ensembles des caractéristiques sélectionnées, le classificateur KNN et une base de données de l'UCI sont utilisés.

Razzaghi T. et al.[125] ont utilisé des techniques de classification : NB, RBFNN, 5-Nearest Neighbours, DT, SVMs et LR pour identifier les complications de la chirurgie bariatrique pour chaque patient. La combinaison des méthodes de classification a permis d'atteindre des mesures de performance élevées.

Dans l'article [126], les auteurs ont proposé un algorithme de sélection de caractéristiques qui emploie l'algorithme modifié de recherche par diffusion stochastique (SDS). Le réseau de neurones, Naïve Bayes et l'arbre de décision ont été utilisés pour la classification. Pour mettre en œuvre cet algorithme, 140 images normales et 130 images anormales ont été utilisées à partir de l'ensemble de données de l'atlas du génome du cancer (TCGA). Pour l'état des données symboliques, l'histologie des poumons a été obtenue sur un maximum de 3 intervalles de temps différents. Les résultats expérimentaux prouvent que la méthode proposée est capable d'atteindre de meilleurs niveaux de performance par rapport aux méthodes existantes telles que la pertinence maximale de la redondance minimale et la sélection de caractéristiques basée sur la corrélation. Cependant, la base de données utilisée est trop petite.

Dans l'article [127], L'objectif de Bayrak T.et Ogul H. est d'introduire un système d'aide à la décision pour les cliniciens permettant de détecter l'apnée du sommeil et l'efficacité du traitement en utilisant les données d'expression génétique. Ils ont présenté deux méthodes distinctes de classification basées sur NB et SVM avec des méthodes de sélection de caractéristiques telles que ReliefF et Chi-Square. La première méthode permet la classification des patients souffrant d'apnée du sommeil (OSAS) tandis que la deuxième tâche consiste à distinguer les patients OSAS non traités et traités. Les gènes les plus significatifs obtenus par des méthodes de sélection de caractéristiques ont été étudiés via la base de données de toxicogénomique comparative pour déterminer l'associativité de la maladie. La base de données GeneMANIA a été utilisée pour présenter un simple réseau d'interaction génétique de ces gènes. La meilleure exactitude de 100 % a été obtenue en utilisant à la fois les prédicteurs et la méthode de sélection des caractéristiques ReliefF dans le diagnostic du OSAS lorsque le nombre de caractéristiques est réduit à 50. La meilleure exactitude de 97,6% a été obtenue en utilisant Naïve Bayes et ReliefF pour distinguer les non-traités et a traité des patients souffrant d'OSAS. Il a été indiqué que vingt et un des cinquante gènes les plus importants sont associés à l'apnée du sommeil. Les vingt gènes apparentés et 123 liens entre les gènes totaux ont été trouvés dans le réseau d'interaction génique obtenu par GeneMANIA. TMSB10, SFN, CA9,

NNMT, SLTM et PITX1 se sont avérés être des gènes hub. hsa-mir-124-3p et hsa-mir-98-5p sont les miARN les plus importants pour le OSAS. Les résultats montrent que l'approche d'apprentissage automatique basée sur des puces à ADN est prometteuse pour introduire des systèmes d'aide à la décision dans le diagnostic de l'apnée du sommeil et dans le traitement des troubles respiratoires du sommeil.

Cette recherche bibliographique nous a permis de déduire que les techniques de l'IA sont largement utilisées pour la prédiction médicale pour plusieurs raisons telles que (traitement d'images, détection précoce, simulations, etc.).

La figure 2.7 synthétise les différentes applications de l'IA en médecine.

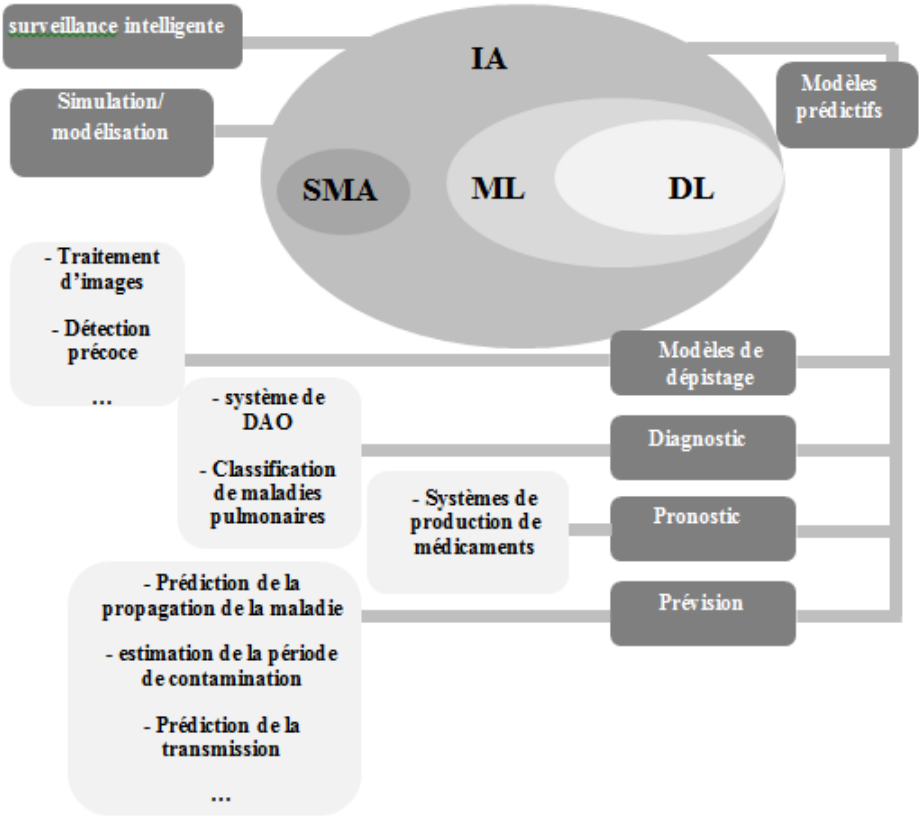


FIGURE 2.7. Application de l'IA pour la prédiction médicale

TABLE 2.4. Résumé des caractéristiques des différents travaux traitant les autres pathologies

Critères / Travaux	Objectif	Technique	Entrées	Mesures de performance									Remarques
				1	2	3	4	5	6	7	8	9	
Boden L.M. et al. 2018 [119]	Prédiction de la probabilité de subir une Chirurgie orthopédique	Méthode mathématique	Rapports cliniques des patients	-	-	-	-	-	✓	✓	-	-	Niveau de preuve faible (4)
Søreide K. et al. 2015 [120]	Prédiction de mortalité pour les patients atteints de maladie gastrique	RNA	Une cohorte de patients ayant subi une chirurgie	-	-	-	-	-	✓	✓	-	-	* Nécessité de l'intervention de l'expert humain pour la collecte des données * Problèmes garbage-in, garbage out
Hunt N. et al. 2018 [121]	Prédiction de la propagation de la rage au Tennessee	analyse spatiale	Rapports du Département de la santé du Tennessee	✓	-	-	-	-	-	-	-	-	* Aide à prévenir les propagations au futur
Devi C.S. et al. 2014 [122]	Diagnostic automatique de maladies cervicales	SMA et ANFIS	Dossiers cliniques des patients	-	-	-	-	-	-	-	-	-	* Diminue les interactions patients/ médecins
Das K. et al. 2016 [123]	Sélection de gènes informatifs	GA , HAS et SVM	BD d'expressions géniques	-	-	-	-	-	-	-	✓	✓	* Efficace pour la suppression des gènes insignifiants et redondants * Réduit le temps de calcul.
Sahebi G. et al. 2017 [124]	Sélection de caractéristiques et optimisation de la classification	GA	UCI Arrhythmia Database	✓	-	-	-	-	-	-	-	✓	* Bonne exactitude * Réduit le temps de calcul.
Razzaghi T. et al. 2017 [125]	Identifier les complications de la chirurgie bariatrique	ML	The Premier Healthcare Database	✓	✓	✓	✓	✓	✓	✓	✓	-	* Bonnes mesures de performance
Shanthi S. et Rajkumar N. 2021 [126]	Prédiction du cancer du poumon	ML	The Premier Healthcare Database	-	✓	✓	-	-	-	-	-	-	* Réduit le temps de calcul
Bayrak T. et Ogul H. 2021 [127]	Prédiction de l'apnée du sommeil	ML +méthode de sélection de caractéristiques	OBI (trois ensembles de données)	✓	-	✓	✓	-	-	-	-	-	* Résultats promoteurs. * Petite BD.

(1 : exactitude, 2 : valeur prédictive positive, 3 : sensibilité, 4 : spécificité, 5 : matrice de confusion, 6 : la courbe ROC, 7 : la zone sous la courbe ROC, 8 : score F1, 9 : temps de calcul)

2.5 Synthèse des travaux

Cette recherche bibliographique traite principalement les maladies considérées comme les principales causes de décès. Les travaux étudiés sont classé selon les pathologies :

1. ***Travaux traitant le cancer du sein***

   Deuxième cause de mortalité chez la femme [128]. Nous avons déduit qu'un grand nombre d'études pour le diagnostic du cancer du sein utilisant des techniques d'apprentissage automatique et en profondeur au cours de la dernière décennie ont fourni des résultats précis. De plus, la plupart d'entre eux ont combiné plusieurs techniques afin d'optimiser les performances et le temps de réponse de leurs systèmes.

   Les réseaux de neurones artificiels (ANN) et la machine à vecteurs de support (SVM) sont les techniques les plus utilisées pour cet intérêt car elles offrent une performance prédictive précise [129].

2. ***Travaux traitant la COVID-19***

   Les technologies de l'IA sont largement utilisées pour suivre, prédire, diagnostiquer et prévoir la pandémie de COVID-19. Nous pouvons remarquer en analysant les travaux rapportés que les techniques de DL sont adaptées pour faire face à la COVID-19. Ainsi, la prise en charge de cette maladie devient plus fiable en réduisant les erreurs de diagnostic et de prédiction, le temps et le coût de calcul.

3. ***Travaux traitant les maladies cardiovasculaires et d'autres maladies courantes telles que les maladies virales, les maladies du système nerveux et d'autres types de cancers***

   Grâce aux techniques de l'IA, la gestion des maladies devient plus fiable en réduisant les erreurs de diagnostic et thérapeutiques, et en extrayant des informations utiles à partir d'une grande quantité de données.

Les tableaux 2.1, 2.2, 2.3 et 2.4 résument les différentes stratégies appliquées à la prédiction médicale. Nous observons que les techniques DL et ML sont largement utilisées pour diagnostiquer et prédire plusieurs pathologies. Nous remarquons également de cette étude bibliographique que la plupart des approches proposées montrent des résultats promoteurs en termes d' exactitude et d'interopérabilité comparés aux résultats obtenus en appliquant les stratégies de base sur des BD référentielles. Cependant, très peu de ces travaux de recherche ont effectivement été intégrés à la

pratique clinique et d'autres travaux de recherche doivent être menés pour la prise en charge efficace du secteur médical.

2.6 L'IA pour la prédiction médicale : enjeux

Malgré les avancées académiques prometteuses apportées par l'application de l'IA dans le domaine médical, cela demeure un challenge et des recherches approfondies supplémentaires sont nécessaires. Nous pouvons résumer ces enjeux en [130] :

1. **Qualité et disponibilité des données :** acquérir de grandes quantités de données cliniques de haute qualité est un processus très difficile, car elles sont dans de multiples formats et fragmentées entre différents systèmes et sources et ont généralement un accès limité.
2. **Problème de sécurité et de confidentialité :** plusieurs chercheurs se sont intéressés à ce concept et ont proposé des travaux pour gérer la sécurité des données [131], car c'est l'un des plus grands défis auxquels sont confrontés les développeurs de systèmes de l'IA. L'exigence de grandes quantités de données provenant de nombreux patients peut affecter la confidentialité des données et leur vie privée.
3. **Problème de biais :** Le biais est l'erreur provenant d'hypothèses erronées dans l'algorithme d'apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues (sous-apprentissage). les systèmes de l'IA apprennent à prendre des décisions sur la base de données d'entraînement pouvant inclure des biais.
4. **Coût de calcul :** la plupart des travaux examinés sont coûteux en termes de calcul, ce qui n'est pas bénéfique pour le clinicien et le patient.
5. **Interprétabilité :** La tâche la plus importante dans le domaine de la santé est d'évaluer et de valider l'approche proposée pour qu'elle soit acceptée par la communauté.
6. **Erreurs prédictives :** un système de l'IA peut parfois se tromper en échouant dans la prédiction des maladies ou dans la recommandation d'un médicament ou dans la prédiction de la réponse d'un patient à un traitement spécifique.

7. **Non-implémentation dans le monde réel :** certains travaux proposés dans la littérature ne sont appliqués dans le monde réel et la cause réside dans la crainte des patients, des médecins et de la société envers l'intégration de l'outil informatique à la médecine.

2.7 Conclusion

La prédiction médicale est un défi pour les cliniciens car elle a une influence directe sur leur pratique quotidienne. Au cours des dernières années, le taux de mortalité a considérablement augmenté, ce qui a nécessité des méthodes et des outils pour une détection précise et précoce des maladies. En étudiant les travaux réalisés dans ce contexte, nous avons remarqué que les chercheurs s'intéressent à la prédiction médicale en utilisant des méthodes et des approches de l'intelligence artificielle. On peut noter que les approches proposées sont efficaces en termes d'exactitude ; cependant, la plupart d'entre elles prennent beaucoup de temps dans la phase d'apprentissage. On peut également remarquer que très peu de ces travaux de recherche ont effectivement été intégrés dans la pratique clinique.

L'application des outils de l'IA dans le domaine médical a atteint un remarquable succès dans l'aide à la décision, le soutien médical des médecins et des patients et la prédiction. Cependant, elle affronte de très grands défis (qualité et disponibilité des données, problème de sécurité et de confidentialité, problème de biais, interprétabilité et erreur prédictive). Pour faire face à ces défis plusieurs solutions sont proposées [130] : Génération et disponibilité de données de qualité, une bonne phase d'apprentissage, une bonne exploitation des méthodes de l'IA (Hybridation des méthodes d'apprentissage profond avec des algorithmes d'optimisation et le parallélisme) qui peut être efficace pour réduction du temps et des coûts.

Dans ce chapitre, nous avons étudié quelques travaux appliquant différentes approches pour la prédiction médicale. Dans le prochain chapitre, nos contributions avec leurs aspects conceptuels et fonctionnels seront présentées.

Chapitre 3

L'IA pour la prédiction médicale : Contributions

3.1 Introduction

L'accès aux innovations thérapeutiques dans la prise en charge du cancer du sein et de la pandémie de la covid-19 sont en progression fulgurante. D'un côté, le cancer du sein touche de plus en plus de femmes à travers le monde. Il se produit lors d'une croissance incontrôlée des cellules dans le tissu mammaire. Le diagnostic du cancer du sein basé sur des données cliniques et histopathologiques peut fournir des résultats incomplets ou trompeurs. Au cours de la dernière décennie, les techniques d'apprentissage automatique (ML) et de l'apprentissage en profondeur (DL) ont été amplement utilisées dans le diagnostic et le pronostic du cancer du sein pour aider les pathologistes dans la détection précoce, la prise de décision et l'élaboration d'un plan de traitement efficace. Dans la littérature, de nombreux algorithmes de classification et de prédiction du cancer du sein sont proposés. D'un autre côté, depuis Décembre 2019, l'apparition de la pandémie Covid-19 a causé un grand nombre de décès dans le monde et cela a touché toutes les tranches d'âge de la population au niveau mondial. Elle a également eu un impact négatif sur le secteur sanitaire, économique, social, etc. Face à cette pandémie, les chercheurs se sont intéressés à l'application des technologies de l'IA pour la prédiction, le diagnostic, et le suivi de cette maladie [132]. Pour ce travail, nous nous sommes intéressés à ces deux maladies et le but de ce chapitre consiste à présenter les principales contributions de cette thèse :

1. Faire une étude comparative de la performance des techniques de ML les plus utilisées pour le diagnostic du cancer du sein.
2. Proposer une approche basée sur la méthode d'extraction de caractéristiques PCA et la régression logistique pour la classification des tumeurs mammaires.
3. Proposer une approche basée sur le DL, les SMA et métaheuristiques pour la prédiction médicale et l'appliquer sur la Covid-19.

3.2 Etude comparative des méthodes de ML pour le diagnostic du cancer du sein

Pour choisir la méthode que nous allons utiliser pour notre première contribution, nous avons effectué une comparaison entre la performance des techniques de ML suivantes : machines à vecteurs de support du noyau (K-SVM), les machines à vecteurs de support linéaires (L-SVM), la régression logistique (LR), les arbres de décision (DT), les k plus proche voisins (k- NN), forêt aléatoire (RF) et le perceptron multicouches (MLP).

3.2.1 Outils et plateformes utilisés

Dans cette étude, La Wisconsin Diagnosis Breast Cancer (WDBC) dataset extraite du référentiel d'apprentissage automatique UCI (tableau A.1) est utilisée [133].
Après avoir normalisé les données d'entrée à l'aide de la méthode z-score, elles ont été divisée en deux sous-ensembles : un sous-ensemble d'apprentissage (75 %) pour entraîner le modèle et un sous-ensemble de test (25 %) pour l'évaluer. Ensuite, une comparaison entre les algorithmes d'apprentissage automatique les plus utilisés dans la littérature est effectuée. Nous avons appliqué des machines à vecteur de support à noyau et linéaire (k-svm et l-svm respectivement), la régression logistique (LR), les arbres de décision (DT), les k plus proches voisins (k-NN), la forêt aléatoire (RF) et le perceptron multicouche (MLP) pour le diagnostic du cancer du sein.

3.2.2 Résultats expérimentaux

Pour évaluer les performances des algorithmes d'apprentissage automatique de cette étude, nous les avons comparés en termes de FPR, TPR, exactitude (apprentissage et test), les instances correctement et incorrectement classées, les matrices de confusion, la courbe ROC et la zone sous la courbe (AUC). Les résultats obtenus sont présentés dans la section suivante.

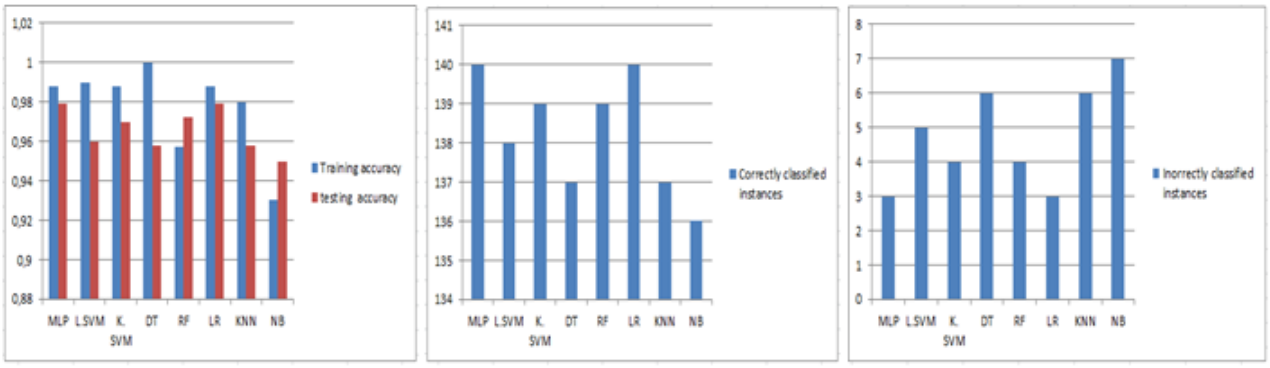
— **instances correctement et incorrectmenet classées, exactitude (apprentissage et test), TPR et FPR**

L'exactitude est une métrique utilisée pour évaluer les modèles de classification. Elle donne la proportion du nombre total de prédictions correctes (Voir équation 1.5).

Le TPR calcule la fraction d'exemples positifs correctement classés. Le FPR calcule la fraction d'exemples négatifs qui sont classés à tort comme positifs (voir Voir équations 1.7 et 1.9).

TABLE 3.1. Evaluation des méthodes.

Evaluation criteria	Classifiers							
	MLP	L-SVM	K-SVM	DT	RF	LR	KNN	NB
Correctly classified instances	140	138	139	137	139	140	137	136
Incorrectly classified instances	3	5	4	6	4	3	6	7
Training accuracy	0,988	0,99	0,988	1	0,957	0,988	0,98	0,93
Testing accuracy	0,979	0,96	0,97	0,958	0,972	0,979	0,958	0,95
TPR	0.963	0.981	0.963	0.926	0.944	0.981	0.944	0.944
FPR	0.011	0.045	0.022	0.022	0.011	0.022	0.034	0.045



(a) Training/ Testing accuracy (b) Correctly classified instances (c) Incorrectly classified instances

FIGURE 3.1. Graphes comparatifs des classificateurs utilisés

— **la courbe ROC**

Pour évaluer l'efficacité des modèles comparés dans cette étude, la courbe ROC a été utilisée (figure 3.2 [134]). C'est un système d'évaluation pour la classification binaire, il permet d'illustrer la précision des classificateurs. La courbe ROC donne des informations sur le taux de vrais positifs (TPR) par rapport au taux de faux positifs (FPR).

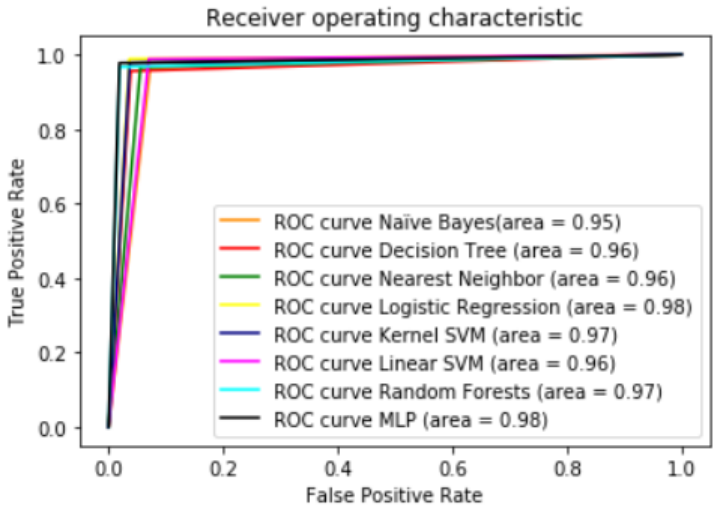


FIGURE 3.2. La courbe ROC

— **les matrices de confusion**

La matrice de confusion fournit des informations sur les classifications réelles et prévues effectuées par un système de classification. Elle est généralement utilisée pour l'évaluation de la performance d'un système (voir chapitre I). Pour comparer les classes réelles et les résultats prédits, nous utilisons les matrices de confusion présentées dans la figure 3.3 [134].

3.2.3 Discussion

En observant les résultats expérimentaux présentés dans le tableau 3.1 et la figure 3.1 [134] nous constatons que l'exactitude obtenue par les méthodes MLP et LR est la meilleure avec une valeur de 98% par rapport à l'exactitude obtenue par KNN, DTs, RF, L-SVM, K-SVM et NB qui varient entre 95% et 97%.

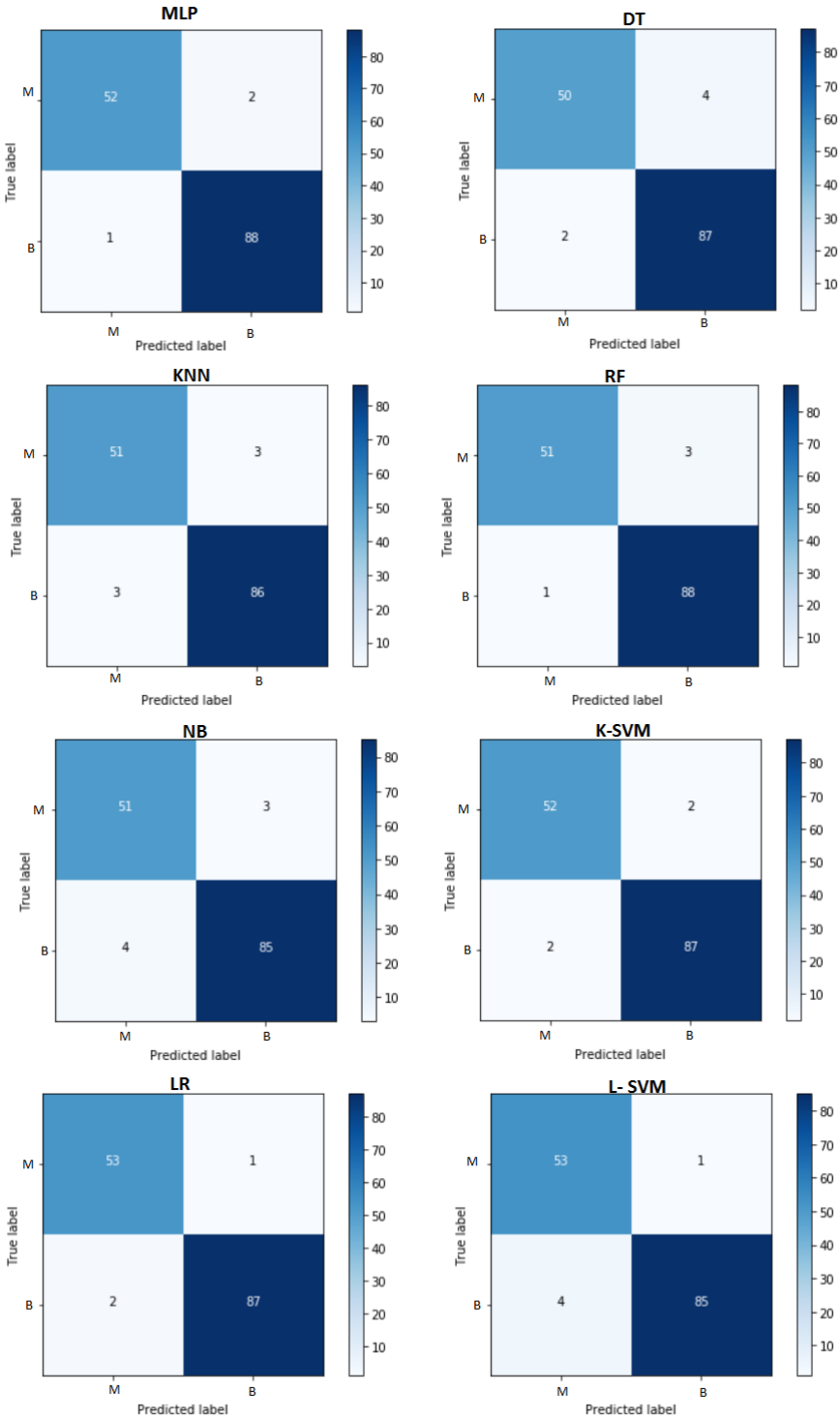


FIGURE 3.3. Les matrices de confusion

Les méthodes MLP et LR atteignent également le meilleur nombre d'instances correctement classées et le nombre minimal d'instances incorrectement classées par rapport aux autres méthodes. De plus, il est à noter que la performance de la méthode DT de la phase d'apprentissage est meilleure que celle de la phase de test ; cela peut être expliqué par le fait que la méthode DT peut apprendre avec exactitude, mais elle peut être faible en généralisation.

La courbe ROC indique la capabilité de distinctions entre les classes (tumeur maligne ou bénigne). La figure 3.2 [134] montre que MLP et LR ont surperformé les autres algorithmes avec une valeur de zone sous la courbe ROC (AUC) de 98%, suivis des algorithmes RF et k-SVM avec une valeur AUC de 97% suivis des algorithmes DTs, KNN et linear SVM avec une valeur de 96% suivis par l'algorithme NB avec une valeur de 95% .

A partir des matrice de confusion présentées dans la figure 3.3 [134], nous notons que MLP et LR prédisent correctement 140 instances sur 143 (87 instances bénignes qui sont effectivement bénignes et 53 instances malignes qui sont malignes), et seulement 3 instances mal prédites (2 instances de classe bénigne prédites comme malignes et 1 instance de classe maligne prédites comme bénignes).

En résumé, les deux algorithmes MLP et LR ont prouvé leur efficacité dans la classification binaires des tumeurs mammaires dans le diagnostic du cancer du sein.

3.3 Un système de diagnostic assisté par ordinateur pour la classification des tumeurs mammaires

Après avoir effectué une étude comparative des méthodes du ML les plus utilisées dans la littérature, un CAD pour la détection du cancer du sein basé PCA-LR est proposé. Dans cette section, l'architecture (figure 3.4) et le processus de l'approche proposée (figure 3.5) seront présentés [135].

3.3.1 Conception du système

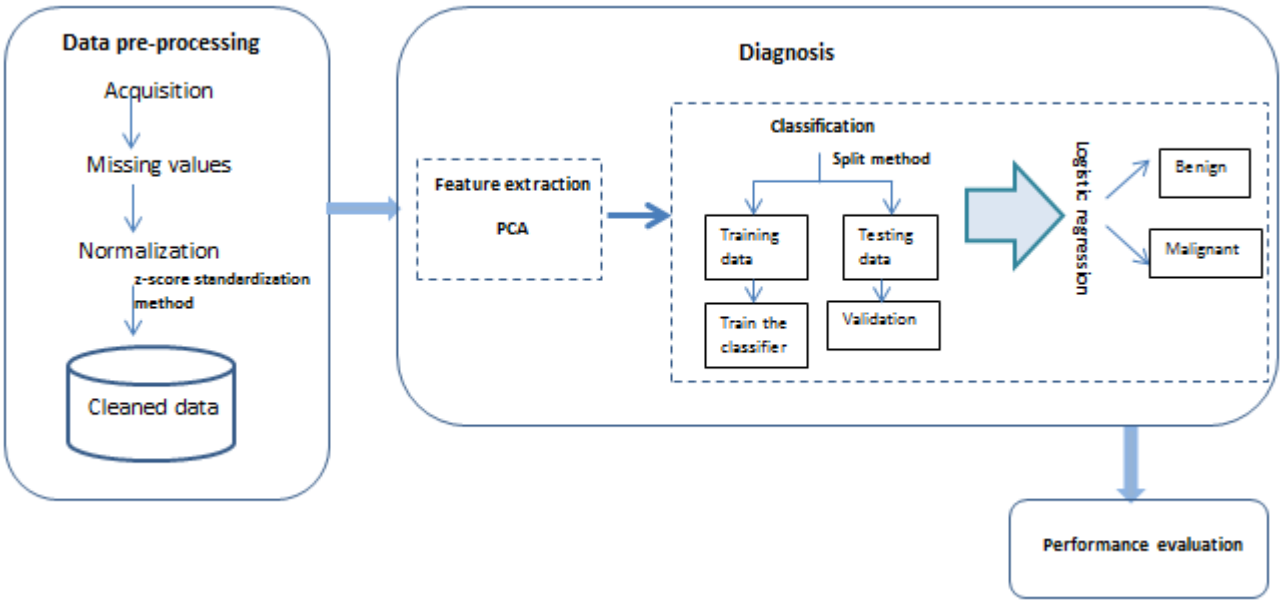


FIGURE 3.4. Une approche basée PCA-LR pour le diagnostic du cancer du sein

La conception du CAD proposé dans cette contribution repose sur trois principaux modules :

1. **Un module de prétraitement des données :**

   Les données médicales sont généralement hybrides, incomplètes, de sources différentes, distribuées sur plusieurs systèmes et leur accès est limité. En conséquence, l'étape de prétraitement des données devient cruciale.

   Dans cette approche, elle consiste en deux étapes :

   (a) Elimination des valeurs manquantes en utilisant KnnImputer de la bibliothèque Scikit-learn.

   Avec cette méthode les valeurs manquantes sont complétées en utilisant k-Nearest Neighbors où, dans chaque échantillon, elles sont imputées avec la valeur moyenne de n_neighbors les plus proches voisins trouvés dans l'ensemble d'apprentissage.

   (b) Normalisation des données en appliquant la méthode de normalisation z-score.

2. **Un module de diagnostic :**

   Après le prétraitement des données, la seconde phase est le diagnostic. Elle est divisée en deux étapes :

(a) Extraction des caractéristiques : Cette étape vise à déterminer les caractéristiques les plus pertinentes de l'ensemble de données pour réduire son volume. Elle a un impact sur la performance du système, la mémoire utilisée et le coût de calcul. Pour ce faire, la méthode PCA pour la réduction de la dimensionnalité des caractéristiques est appliquée. Son idée principale est de transformer les variables corrélées en de nouvelles variables appelées composantes principales. Le pseudo-code de l'algorithme PCA est donné dans l'algorithme 1 [135].

Algorithme 1 : Algorithme PCA

Input : Normalized dataset X with size N × M ▷ $X_i = (x_{1i}, x_{2i}, \ldots, x_{Mi})$, $i = 1, 2, 3, \ldots, N$

Output : Reduced data with size N × K

1: Compute the mean of each column, putting it into matrix **B**. $\mu \leftarrow \frac{1}{N}\sum_{i=1}^{N} X_i$

2: Compute covariance matrix of the dataset $\mathbf{C} \leftarrow \frac{1}{N} B^T B.$

3: Compute the Eigen values (λ_j) and Eigen vectors (v_j) of **C**, $Cv_j = \lambda_j v_j$, $j = 1, 2, 3, \ldots, M$

4: Estimate high-valued Eigen vectors

(i) Choose a threshold θ

(ii) Select K Eigen vectors corresponding to selected high-valued λ_j ▷ Reject those with Eigen value less than θ

5: Reduce the high dimensionality of feature matrix from M to K

(b) La classification des tumeurs mammaires en appliquant la méthode de régression logistique. Le processus de cette approche est détaillé dans le diagramme de séquence (figure figure 3.5 et l'algorithme 2 [135]).

Algorithme 2 : Une approche basée PCA-LR pour le diagnostic du cancer du sein

Input : WBCD dataset

Output : Y = tumor class (benign or malignant)

1. Data acquisition ;
2. Missing values imputation ;
3. Normalization ▷ Equation 1
4. Feature extraction ▷ Algorithm 2
5. Data split

X_train, Y_train, X_test, Y_test = split(x,y) ▷ 70% training, 30% testing

6. Classifier training ▷ Logistic regression
7. Model evaluation

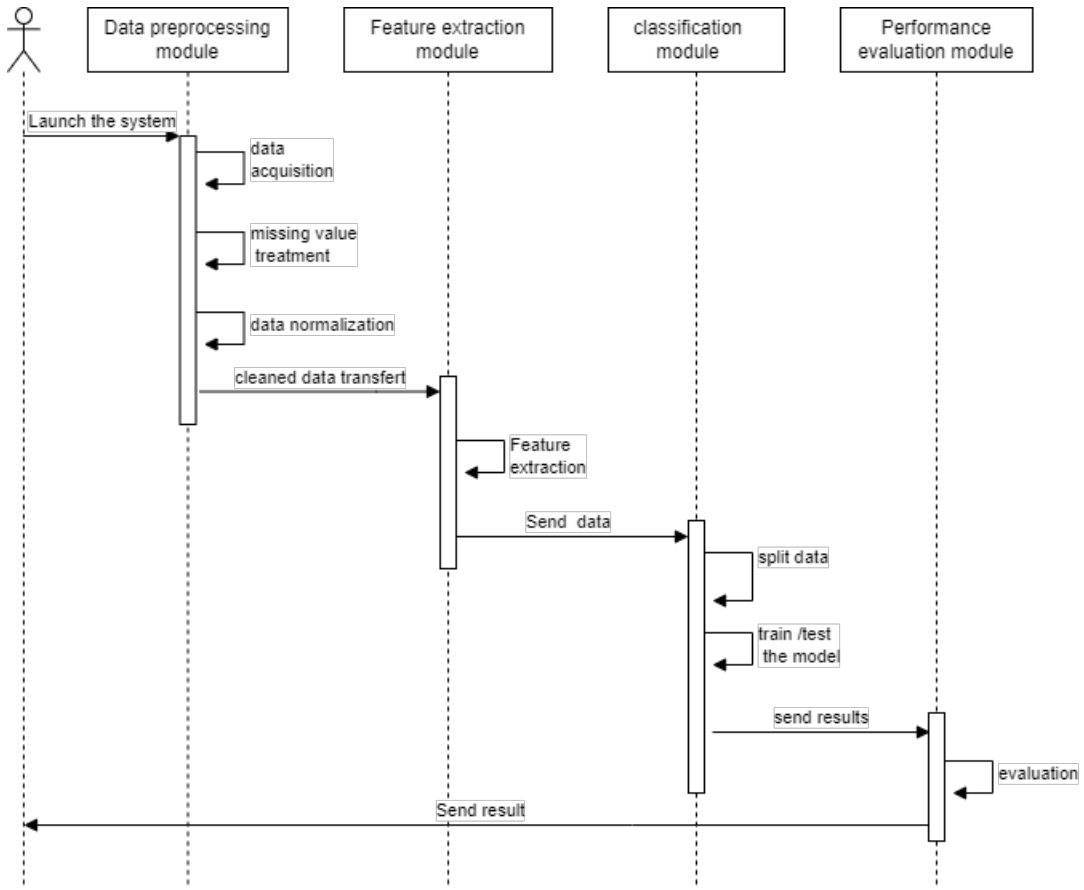


FIGURE 3.5. Diagramme de séquence de l'approche basée PCA-LR pour le diagnostic du cancer du sein

3. **Un module d'évaluation**

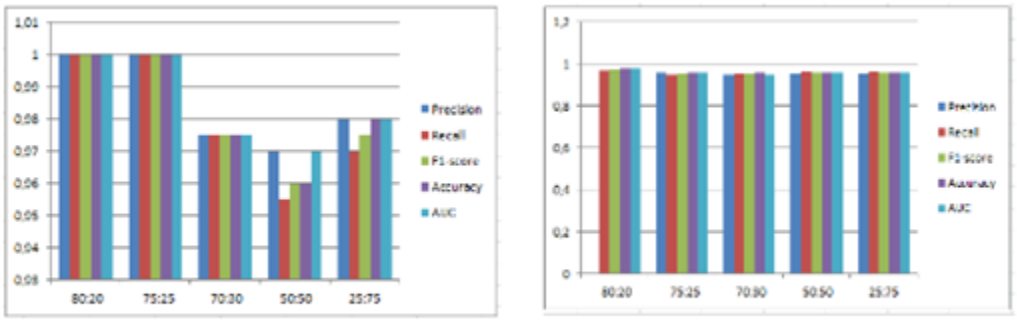
   Pour évaluer et valider l'approche proposée, elle a été comparée à d'autres méthodes d'extraction et de sélection de caractéristiques combinées avec la régression logistique, elle a également été comparée à d'autres méthodes de classification en termes de précision (PPV), de rappel (sensibilité), de score F1, d'exactitude, de zone sous la courbe ROC (AUC), de courbe ROC et de matrice de confusion. Les expérimentations et les résultats sont présentés dans la section ci-dessous.

3.3.2 Résultats expérimentaux et discussion

Les expérimentations ont été réalisées sur les deux bases de données publiques WDBC [133] et WOBC [136] (tableaux A.1 et A.2). Nous pouvons les résumer comme suit :

1. **Expérimentation 1 : Application du modèle proposé sur différentes partitions d'entraînement-test**

   La taille du sous-ensemble d'apprentissage et du sous-ensemble de test peut avoir un impact sur la performance du modèle de classification. Pour cela, la première expérimentation consiste à comparer la performance du modèle proposé sur différentes partitions d'entraînement-test : 80-20 %, 75-25 %, 70-30 %, 50-50 % et 25-75 %. Les résultats expérimentaux sont présentés dans le tableau 3.2 et la figure 3.6 [135].



(a) Avec WDBC (b) Avec WOBC

FIGURE 3.6. Résultats obtenus en utilisant différentes partitions d'apprentissage-test

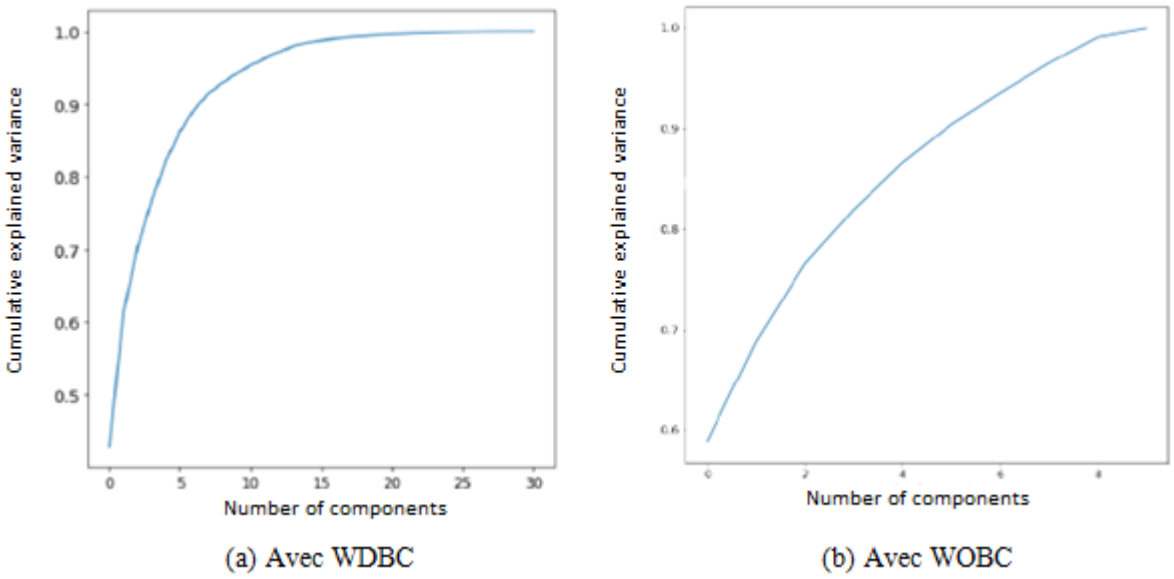
Pour l'ensemble de données WOBC, les meilleurs résultats en termes d'exactitude, de rappel, de score F1, de précision et d'AUC avec des valeurs de (0,98, 0,97, 0,975, 0,98 et 0,98) sont obtenus avec les partitions 80-20 %. Concernant l'ensemble de données WDBC, le résultat obtenu est 1 pour toutes les métriques pour les partitions 80-20 %, 75-25 %. Ces résultats montrent que l'utilisation d'une grande partition d'entrainement permet un meilleur apprentissage et une bonne généralisation.

TABLE 3.2. Résultats obtenus en utilisant différentes partitions d'apprentissage-test

Dataset	Train : Test ratio	Classe	PPV	Rappel	score F1	Accuracy	AUC
WDBC	80 : 20	M	1.00	1.00	1.00	1.00	1.00
		B	1.00	1.00	1.00		
		Mean	1.00	1.00	1.00		
	75 : 25	M	1.00	1.00	1.00	1.00	1.00
		B	1.00	1.00	1.00		
		Mean	1.00	1.00	1.00		
	70-30%	M	0.97	0.97	0.97	0.98	0.97
		B	0.98	0.98	0.98		
		Mean	0.975	0.975	0.975		
	50 : 50	M	0.98	0.92	0.95	0.96	0.97
		B	0.96	0.99	0.97		
		Mean	0.97	0.955	0.96		
	25-75%	M	1.00	0.94	0.97	0.98	0.98
		B	0.96	1.00	0.98		
		Mean	0.98	0.97	0.975		
WOBC	80-20%	M	0.98	0.96	0.97	0.98	0.98
		B	0.98	0.99	0.98		
		Mean	0.98	0.97	0.975		
	75-25%	M	0.95	0.93	0.94	0.96	0.96
		B	0.97	0.97	0.97		
		Mean	0.96	0.95	0.955		
	70-30%	M	0.93	0.95	0.94	0.96	0.95
		B	0.97	0.96	0.97		
		Mean	0.95	0.955	0.955		
	50-50%	M	0.93	0.97	0.95	0.96	0.96
		B	0.98	0.96	0.97		
		Mean	0.955	0.965	0.96		
	25-75%	M	0.93	0.96	0.95	0.96	0.96
		B	0.98	0.97	0.97		
		Mean	0.955	0.965	0.96		

2. **Expérimentation 2 : Application sur différents sous-ensembles de caractéristiques**

   Pour vérifier l'effet de la réduction de la dimensionnalité sur l'efficacité de l'approche proposée, nous l'avons testée sur divers sous-ensembles de caractéristiques (85 %, 90 %, 95 %, 97 % et 99 %) sur les ensembles de données WDBC et WOBC.



(a) Avec WDBC (b) Avec WOBC

FIGURE 3.7. Résultats obtenus en utilisant différents sous-ensembles de caractéristiques

Pour cette étude, 95 % et 85 % de variance expliquée et cumulée pour WDBC et WOBC ont été sélectionnées par PCA (voir figure 3.7 [135]).

Pour WDBC, en réduisant le nombre original de caractéristiques à 11, le système atteint sa meilleure performance avec une PPV, un rappel, un score F1, une exactitude et une AUC de 1,00. Pour WOBC, avec différentes variances, les performances du système sont constantes pour toutes les métriques avec une PPV de 0,98, un rappel de 0,97, un score F1 de 0,98, une exactitude de 0,98 et une AUC de 0,98 (voir tableau 3.3 [135]).

Les caractéristiques sélectionnées sont les plus pertinentes et permettent d'améliorer les mesures de performance et de réduire la consommation de temps et le coût de calcul.

TABLE 3.3. Résultats obtenus avec une variance de 85% 90%, 95%, 97% et 99% of variance

Dataset	% variance	Nombre de caractéristiques sélectionnées	PPV	Rappel	Score F1	exactitude	AUC
WDBC	85%	6	0.99	0.99	0.99	0.99	0.99
	90%	8	0.97	0.97	0.97	0.97	0.97
	95%	11	1.00	1.00	1.00	1.00	1.00
	97%	13	1.00	1.00	1.00	1.00	1.00
	99%	18	1.00	1.00	1.00	1.00	1.00
WOBC	85%	5	0.98	0.97	0.98	0.98	0.98
	90%	6	0.98	0.97	0.98	0.98	0.98
	95%	8	0.98	0.97	0.98	0.98	0.98
	97%	9	0.98	0.97	0.98	0.98	0.98
	99%	9	0.98	0.97	0.98	0.98	0.98

3. **Expérimentation 3 : Comparaison de PCA à d'autres méthodes de réduction et de sélection des caractéristiques (ISOMAP, algorithme Relief)**

   Dans la littérature, plusieurs méthodes de réduction de la dimensionnalité et de sélection des caractéristiques sont proposées. Pour évaluer l'efficacité de la combinaison des méthodes PCA et LR pour le diagnostic du cancer du sein, nous avons effectué une comparaison de la méthode PCA avec les méthodes ISOMAP et l'algorithme Relief. Le tableau 3.4 rapporte les résultats expérimentaux de cette comparaison [135].

   Nous pouvons remarquer que PCA surperforme les autres techniques en terme de PPV, de rappel, de score F1, d'exactitude et d'AUC de valeurs (0,98, 0,97, 0,98, 0,98, et 0,98) sur WOBC et une valeur de 1 pour toutes les métriques sur l'ensemble de données WDBC. En combinant ISOMAP et l'algorithme Relief avec LR, la meilleure sélection est 6 et 9 fonctionnalités sur WDBC. Ainsi, un nombre réduit de caractéristiques permet de minimiser le temps de calcul. Cependant, nous remarquons que les performances de classification sont inférieures à celles de l'approche proposée qui sélectionne 11 caractéristiques. Cela indique que le modèle proposé a atteint la plus grande exactitude de classification en améliorant la

qualité des données, en diminuant le nombre d'attributs sans perdre les principales informations objectives. Comme indiqué dans le tableau, pour l'ensemble de données WOBC, l'algorithme Relief et PCA fournissent des résultats identiques. Cependant, PCA surpasse l'algorithme Relief pour les deux ensembles de données et assure la capacité de généralisation.

TABLE 3.4. Comparaison de la méthode PCA avec d'autres méthodes de réduction

Dataset	Méthode	nombre de caractéristiques seléctionnées	Classe	PPV	Rappel	Score F1	Exactitude	AUC
WDBC	ISOMAP + LR	6	M	0.96	0.98	0.97	0.98	0.98
			B	0.99	0.98	0.98		
			Moyenne	0.98	0.98	0.975		
	Relief Algo + LR	9	M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Moyenne	0.98	0.98	0.98		
	Approche proposée (PCA + LR)	11	M	1.00	1.00	1.00	1.00	1.00
			B	1.00	1.00	1.00		
			Moyenne	1.00	1.00	1.00		
WOBC	ISOMAP + LR	5	M	0.98	0.93	0.95	0.97	0.97
			B	0.97	0.99	0.98		
			Moyenne	0.97	0.96	0.97		
	Relief Algo + LR	5	M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Moyenne	0.98	0.97	0.98		
	Approche proposée (PCA + LR)	5	M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Moyenne	0.98	0.97	0.98		

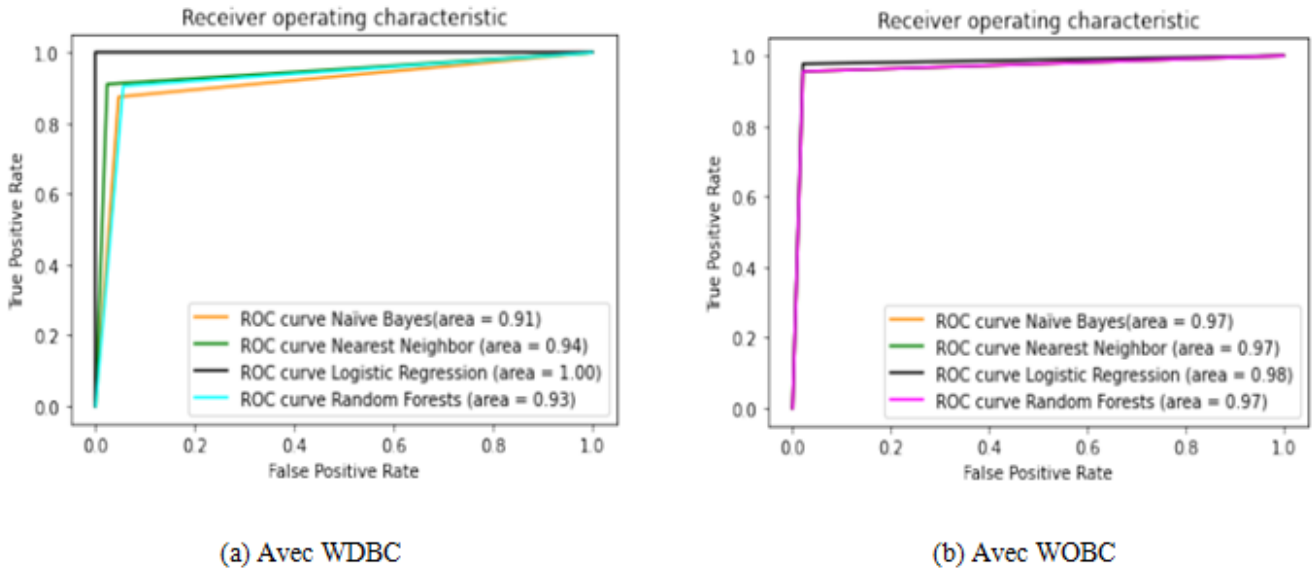
4. **Expérimentation 4 : Comparaison de la LR avec d'autres méthodes de ML (NB, RF, Knn)**

Pour évaluer l'efficacité de régression logistique sur le diagnostic du cancer du sein, une comparaison de la méthode proposée avec les autres méthodes classiques de ML, à savoir, NB, RF, Knn est effectuée.

TABLE 3.5. Comparaison de l'approche proposée avec d'autres méthodes de ML

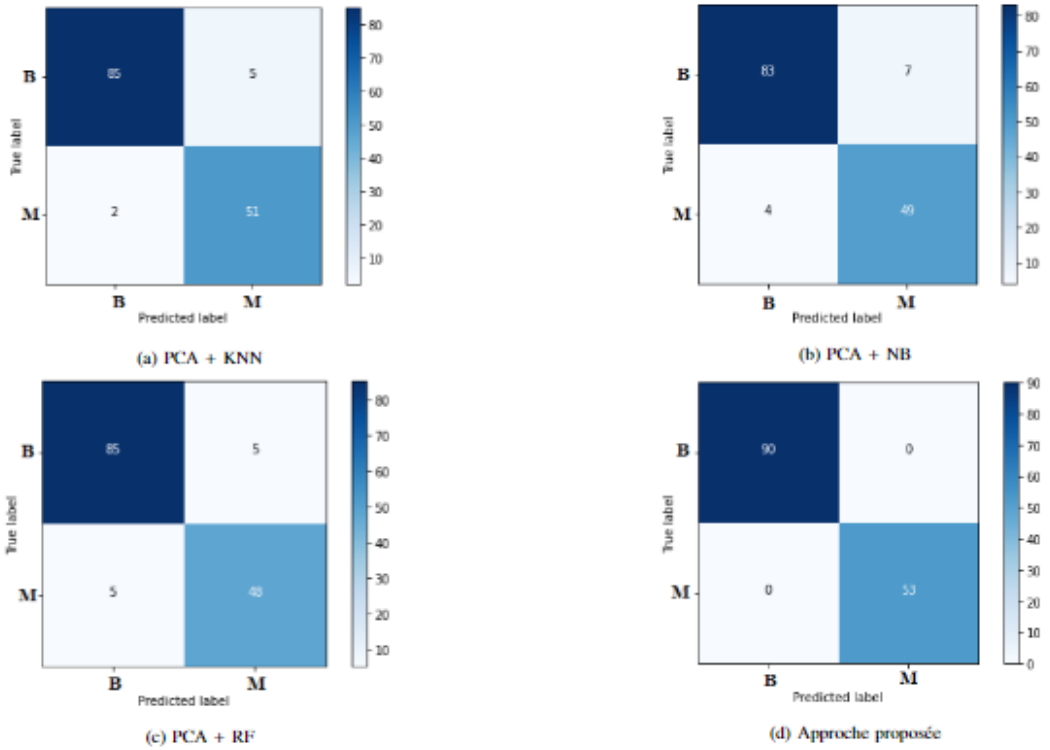
Dataset	Méthode	classe	PPV	Rappel	Score F1	Exactitude	AUC
WDBC	PCA +NB	M	0.88	0.92	0.90	0.92	0.91
		B	0.95	0.92	0.94		
		Moyenne	0.915	0.92	0.92		
	PCA + RF	M	0.91	0.91	0.91	0.93	0.93
		B	0.94	0.94	0.94		
		Moyenne	0.925	0.925	0.925		
	PCA + Knn	M	0.91	0.96	0.94	0.94	0.95
		B	0.98	0.94	0.96		
		Moyenne	0.945	0.95	0.95		
	approche proposée PCA + LR	M	1.00	1.00	1.00	1.00	1.00
		B	1.00	1.00	1.00		
		Moyenne	1.00	1.00	1.00		
WOBC	PCA +NB	M	0.96	0.96	0.96	0.97	0.97
		B	0.98	0.98	0.98		
		Moyenne	0.97	0.97	0.97		
	PCA + RF	M	0.96	0.96	0.96	0.97	0.97
		B	0.98	0.98	0.98		
		Moyenne	0.97	0.97	0.97		
	PCA + Knn	M	0.98	0.98	0.98	0.97	0.97
		B	0.96	0.96	0.96		
		Moyenne	0.97	0.97	0.97		
	approche proposée PCA + LR	M	0.98	0.96	0.97	0.98	0.98
		B	0.98	0.99	0.98		
		Moyenne	0.98	0.97	0.98		

En analysant les résultats du tableau 3.5 et les courbes ROC de la figure 3.8, nous pouvons déduire que la combinaison de la méthode de réduction de la dimensionnalité PCA et la méthode de classification LR est promemeteuse pour le diagostic du cancer du sein[135].



(a) Avec WDBC (b) Avec WOBC

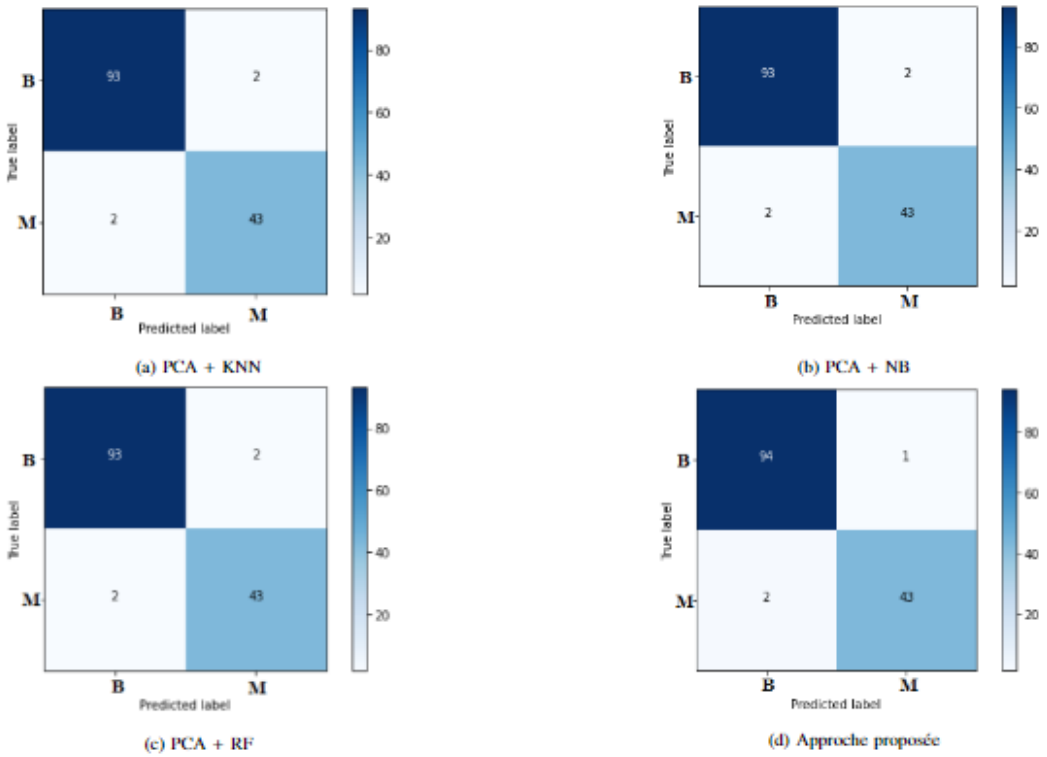
FIGURE 3.8. Courbes ROC



(a) PCA + KNN (b) PCA + NB

(c) PCA + RF (d) Approche proposée

FIGURE 3.9. Matrices de confusion obtenues avec WDBC



(a) PCA + KNN

(b) PCA + NB

(c) PCA + RF

(d) Approche proposée

FIGURE 3.10. Matrices de confusion obtenues avec WOBC

les matrices de confusion (figures 3.9 et 3.10) montrent qu'avec l'approche proposée aucune instance mal classées n'est obtenue alors que les autres méthodes (KNN, NB et RF) prédisent respectivement 7, 11 et 10 instances de manière incorrect sur l'ensemble de données WDBC. Concernant l'ensemble de données WOBC, nous remarquons que l'approche proposée apporte une légère amélioration par rapport aux autres méthodes en prédisant incorrectement 3 instances là où les autres techniques échouent sur 4 prédictions[135].

5. **Expérimentation 5 : Comparaison de l'approche proposée avec d'autres approches de la littérature**

   Le tableau 3.6 présente les résultats d'une comparaison entre l'approche proposée et des études existantes dans la littérature dans différentes conditions[135]. Les études précédentes sont basés sur différentes techniques ML telles que MPL, RF, LR, LSSVM, ANN et PCA, BN et RBF pour le diagnostic du cancer du sein.

   Les résultats obtenus confirment que la combinaison de PCA et la LR est très prometteuses sur l'ensemble de données WDBC en termes de précision (PPV), de rappel, de score F1,

d'exactitude et d'AUC avec une valeur de 1,00. De plus, sur l'ensemble de données WOBC, elle apporte une légère amélioration en termes de précision (PPV), d'exactitude, de score F1 et d'AUC et donne les meilleures valeurs de rappel par rapport aux autres études. Pour assurer le diagnostic précoce du cancer du sein et éviter ses complications, on s'intéresse à prédire les tumeurs malignes (les cas positifs). Par conséquent, le rappel est considéré comme une mesure importante puisqu'il indique le taux d'échantillons malins correctement identifiés.

TABLE 3.6. Comparaison de l'approche proposée avec d'autres approches de la littérature

Dataset	méthode de Classification	Train : Test ratio	Classe	PPV	Rappel	Score F1	Exactitude	AUC
WDBC	MLP [134]	75-25%	M	0.99	0.97	0.98	0.98	0.98
			B	0.95	0.98	0.97		
			Moyenne	0.97	0.975	0.975		
	RF [134]		M	0.96	0.99	0.97	0.96	0.97
			B	0.98	0.92	0.95		
			Moyenne	0.97	0.955	0.96		
	LR [134]		M	0.99	0.98	0.99	0.98	0.98
			B	0.97	0.98	0.98		
			Moyenne	0.98	0.98	0.985		
	Approche proposée PCA + LR		M	1.00	1.00	1.00	1.00	1.00
			B	1.00	1.00	1.00		
			Moyenne	1.00	1.00	1.00		
WOBC	PCA + ANN [137]	80-20 %	M	-			0.97	-
			B	-				
			Moyenne	0.95	0.95	0.95		
	LSSVM classifier [138]		M	-			0.97	-
			B	-				
			Moyenne	-	0.97	-		
	Approche proposée PCA + LR		M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Moyenne	0.98	0.97	0.975		
	BN+RBF [84]	75-25%	M	-			0.97	-
			B	-				
			Moyenne	0.993	0.97	0.98		
	Approche proposée PCA + LR		M	0.95	0.95	0.95	0.97	0.96
			B	0.97	0.97	0.97		
			Moyenne	0.96	0.96	0.96		
	LSSVM classifier [138]	50-50%	M	-			0.958	-
			B	-				
			Moyenne	-	0.948	-		
	Approche proposée PCA + LR		M	0.93	0.96	0.94	0.96	0.95
			B	0.98	0.96	0.97		
			Moyenne	0.95	0.96	0.96		

6. **Expérimentation 6 : Comparaison de l'approche proposée avec des modèles basés sur DL**

   La dernière expérimentation consiste à comparer l'approche proposée avec des modèles de diagnostic basés sur le DL en terme d'exactitude. Les résultats (figure 3.11 [135]) montrent que la combinaison de PCA et LR surperforme les autres travaux basés sur le DL avec une exactitude de 100 % sur l'ensemble de données WDBC. Cela peut s'expliquer par le fait que :

   — PCA améliore les performances de la régression logistique et surmonte le problème de surajustement.

   — Les méthodes DL peuvent être inefficaces lorsque les bases de données sont petites.

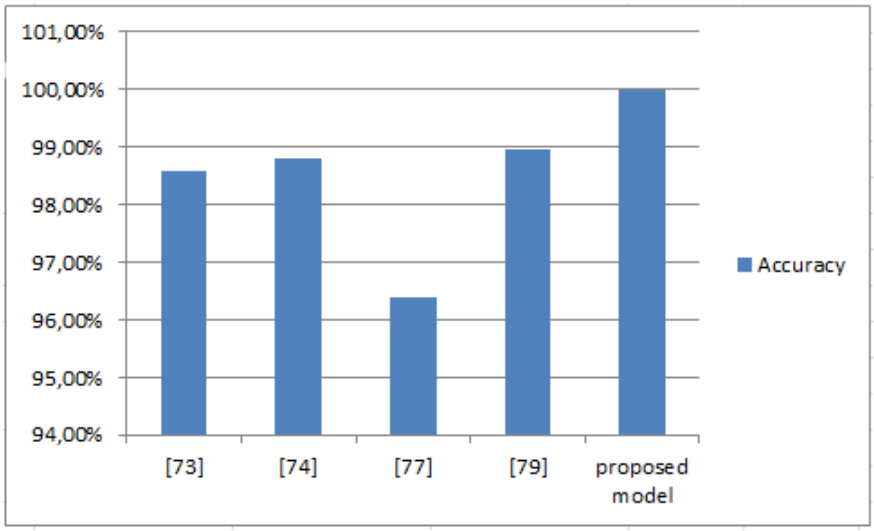


FIGURE 3.11. Comparaison de l'approche proposée avec des modèles basés sur DL

Les résultats promotteurs obtenus sur deux différents ensembles de données et dans différentes conditions prouvent l'efficacité de la combinaison de la méthode d'extraction de caractéristiques PCA avec la méthode de classification LR pour le diagnostic du cancer du sein. Cependant, l'approche proposée n'a pas prouvé son efficacité sur d'autres bases de données plus volumineuses.

3.4 Un système intelligent pour la prédiction de la COVID-19

Le coronavirus 2 du syndrome respiratoire aigu sévère (SARS-CoV-2) est le virus responsable de la maladie coronavirus 2019 (COVID-19). Elle a été signalée pour la première fois à Wuhan

(Chine) fin décembre 2019. Cette pandémie est devenue une menace sanitaire mondiale et elle se propage désormais dans le monde entier, les données statistiques en révélant le pourcentage de décès et de contaminations montrent la gravité de la situation. Pour faire face à cette propagation exponentielle, l'adoption rapide d'outils de l'IA est cruciale. Pour cette raison, de nombreuses recherches en IA sont menées et plusieurs articles de recherche sur ce sujet sont publiés.

Dans cette section, une nouvelle approche basée sur les SMA, DL et les métaheuristiques pour la prise en charge de la COVID-19 sera présentée.

3.4.1 La base de données utilisée

Avec l'explosion du nombre de décès à cause de la Covid-19, beaucoup de chercheurs dans différents domaines se sont intéressés à cette maladie et plusieurs bases de données ont été créées et publiées en ligne. Dans cette section la base de données choisie pour le développement du système proposé sera présentée.

Nous avons collecté une base de données publique « COVID-19 patient pre-condition dataset » de Kaggle [139]. Elle contient différentes informations, elle est composée de 23 attributs et 566602 enregistrements résumés dans le tableau A.3.

La matrice de corrélation (figure 3.12) fournit un excellent visuel lors de la comparaison de plusieurs variables et des relations entre elles.

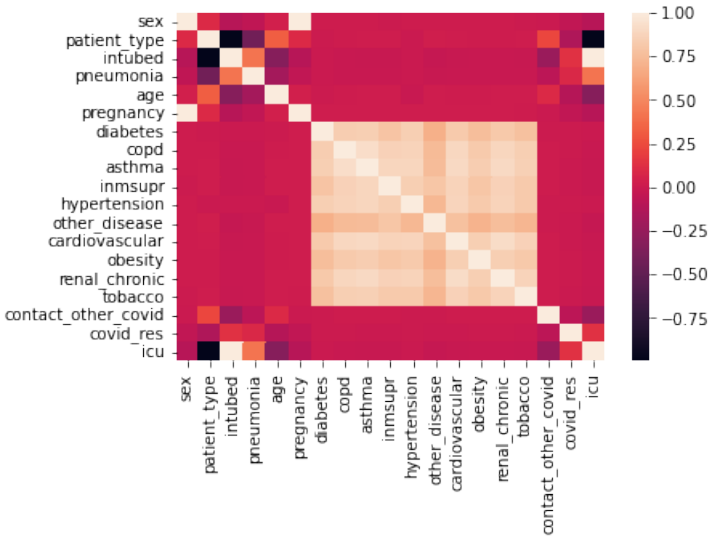


FIGURE 3.12. Matrice de corrélation

3.4.2 Description de l'environnement des agents

La spécification de l'environnement des agents est la première étape pour concevoir un agent [5]. Un environnement se compose d'un ensemble d'agents et des propriétés qui le décrivent. On parle de la description PEAS (Performance, Environnement, Actions et Sensors) comme illustré dans le tableau 3.7 :

- **Preformance :** est le résultat souhaité de l'agent.
- **Environnement :** représentant l'ensemble des conditions et des objets utilisés par l'agent pour accomplir ses tâches.
- **Actions :** outils matériels et logiciels utilisés par l'agent pour effectuer des actions afin de produire un résultat.
- **Sensors :** ensemble de dispositifs utilisés par l'agent pour la perception de son environnement.

TABLE 3.7. Description PEAS

	Mesures de performance	Environnement	Actions	Sensors
Agents pour prédiction médicale	- Précision ; - Temps de réponse ; - Coût ; - Sécurité.	- Patients ; - Médecins ; - Staff médical.	- Questions ; - Diagnostic ; - Cytologie ; - Rapports médicaux.	- Réponses des patients ; - Capteurs ; - GPS ...

3.4.3 Description des caractéristiques de l'environnement

Un environnement est caractérisé par un ensemble de dimensions qui déterminent la conception et la mise en œuvre appropriées de l'agent [140]. Ces dimensions sont résumées dans le tableau 3.8.

TABLE 3.8. Description des propriétés de l'environnement

Entièrement observable Des capteurs donnent accès à l'état complet de l'environnement (aucune donnée manquante).	**Partiellement observable** Données manquantes à cause de capteurs bruyants et imprécis.
Déterministe L'état suivant est complètement déterminé par l'état actuel et les actions de l'agent.	**Stochastique** La sortie n'est pas entièrement déterminée par les conditions initiales.
Episodique Le comportement de l'agent est divisé en épisodes indépendants.	**Séquentiel** La décision actuelle peut affecter les décisions futures.
Statique L'environnement ne change pas avec le temps.	**Dynamique** L'environnement change avec le temps.
Discret Le nombre d'états distincts est limité.	**Continu** L'état change avec le temps.
Agent unique Un seul agent pour résoudre un problème.	**Multi agent** Plusieurs agents pour la résolution du problème (comportement compétitif ou coopératif).

La description des caractéristiques de l'environnement du SMA proposé est présentée dans le tableau 3.9.

TABLE 3.9. Description des propriétés de l'environnement de l'approche proposée

Environnement	Observable	Déterministe	Episodique	Statique	Discret	Agent
	Partiellement	Stochastique	Séquentiel	Dynamique	Continu	Multi (coopératif)

3.4.4 Conception et fonctionnement du système proposé

Pour concevoir notre architecture, nous avons opté pour le paradigme agent pour tirer parti de ses caractéristiques (autonomie, communication, traitement parallèle et collaboration), cela permet de résoudre les défis de l'application des techniques d'IA dans le secteur de la santé pour la gestion de la pandémie de Covid-19 (cités dans le chapitre II).

La conception du système proposé est basée sur l'interaction entre un ensemble d'agents cognitifs et réactifs (figure 3.13.)

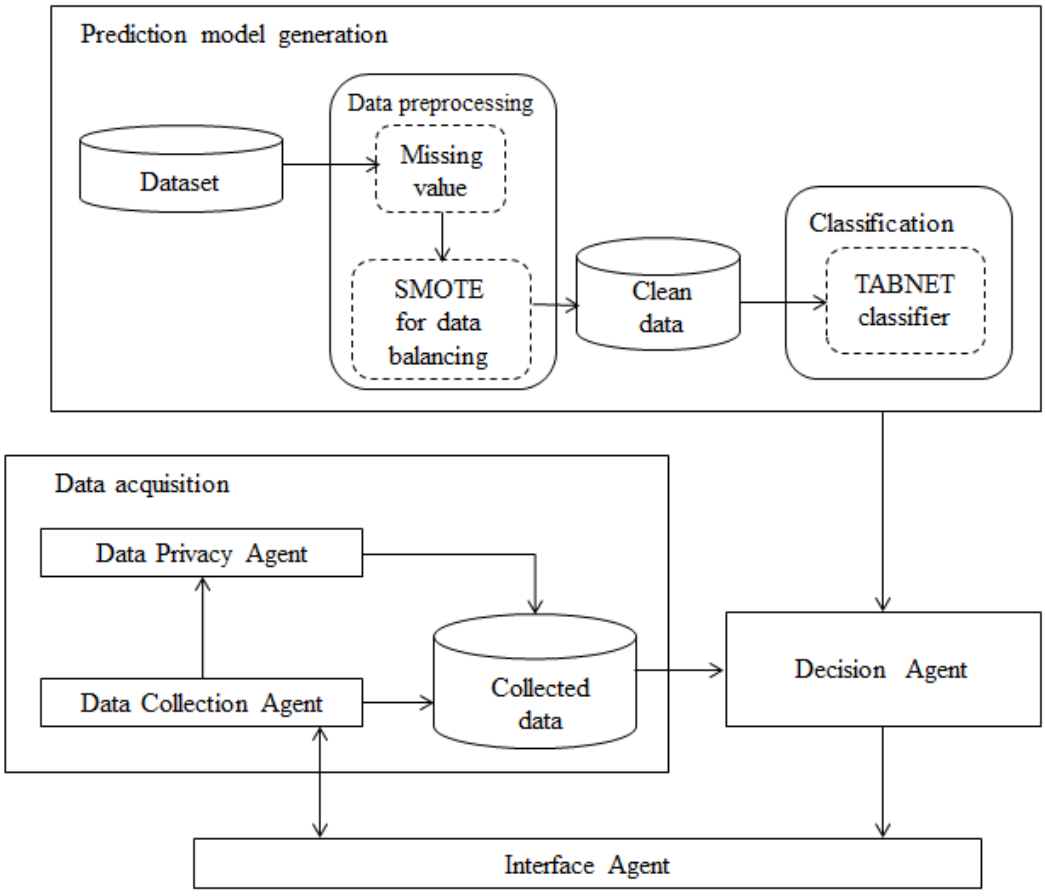


FIGURE 3.13. Une approche basée SMA, DL et méta-heuristique pour la prédiction médicale

3.4.4.1 Génération du modèle de prédiction

Cette phase a pour but la création du modèle de prédiction (résumée dans le diagramme de séquence de la figure 3.16) :

1. **Prétraitement des données**

   C'est une étape cruciale pour le développement d'outils de DL, elle permet d'améliorer la qualité des données d'entrée ce qui contribue à obtenir des résultats plus précis. Les sections suivantes traitent les différentes étapes de la transformation des données d'entrée de leur état brute à un état compréhensible pour l'analyse.

— **Valeurs manquantes**

Cette étape consiste à éliminer ou remplacer les lignes ou colonnes contenant des valeurs manquantes.

Pour visualiser les données manquantes de la base de données utilisée, nous avons tracé l'histogramme (figure 3.14) et nous avons calculé le nombre total pour chaque caractéristique (tableau 3.10.)

TABLE 3.10. Nombre total des données manquantes

Attributs	7	8	10	11	12	13	14	15	16	17	18	19	20	21	23
Nombre total des données manquantes	444813	11	288699	1981	1749	1752	1980	1824	2598	1822	1781	1792	1907	175031	444814

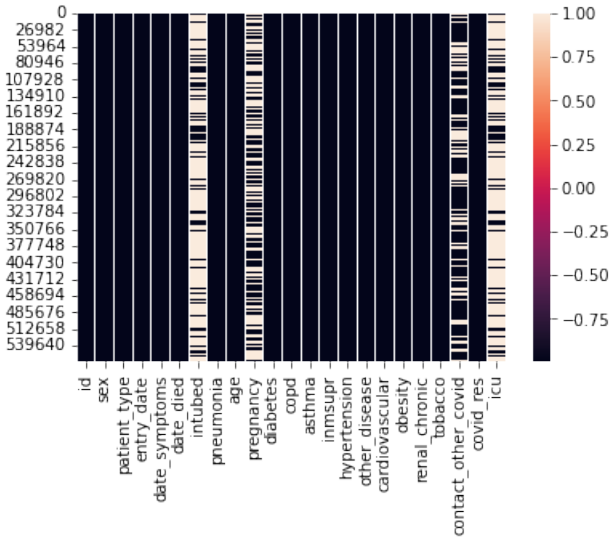


FIGURE 3.14. Histogramme des données manquantes

Pour cette contribution, nous avons opté pour le KNNImputer. Il utilise la méthode k-NN pour remplacer les valeurs manquantes dans les ensembles de données par la valeur moyenne des voisins les plus proches trouvés dans l'ensemble d'apprentissage.

— **Déséquilibre des données**

La distribution déséquilibrée des données se produit généralement lorsque les observations dans l'une des classes sont beaucoup plus élevées ou plus basses que les autres classes. Dans le domaine médical, ce déséquilibre affecte l'exactitude de la classification des diagnostics médicaux.

Pour pallier ce problème, plusieurs méthodes d'échantillonnage consistant à éliminer des échantillons de la classe majoritaire (sous-échantillonnage) et/ou à ajouter plus d'exemples de la classe minoritaire (sur-échantillonnage) sont proposées. Pour cette contribution nous avons utilisé SMOTE (Synthetic Minority Oversampling Technique). SMOTE est une approche de sur-échantillonnage dans laquelle la classe minoritaire est sur-échantillonnée en créant des exemples « synthétiques ». La classe minoritaire est sur-échantillonnée en prenant chaque échantillon de la classe minoritaire et en introduisant des exemples synthétiques le long des segments de ligne joignant les k voisins les plus proches de la classe minoritaire. En fonction de la quantité de sur-échantillonnage requise, les voisins parmi les k voisins les plus proches sont choisis au hasard. Cette méthode est semblable à la data augmentation que l'on utilise pour réduire les risques d'overfitting.

2. **Classification**

   Pour assurer l'étape de la classification, nous avons opté pour le classificateur **TabNet (Attentive Interpretable Tabular Learning)**.

   Proposée en 2019 par un groupe de chercheurs de Google, TabNet est une méthode de classification et de régression basée sur le DL pour l'apprentissage sur des données tabulaires brutes (non normalisées).

   L'architecture TabNet est une succession d'étapes (figure 3.15) [141] :

   - **Feature transformer** qui est un bloc de décision de quatre GLU consécutifs.
   - **Attentative Transformer** qui utilise une matrice clairsemée pour donner une sélection de fonctionnalités clairsemées qui permet une interprétabilité et un meilleur apprentissage car la capacité est utilisée pour les fonctionnalités les plus saillantes.
   - **Mask** qui est utilisé avec le transformateur pour donner le paramètre de décisions : n(d) et n(a) qui est ensuite transmis à l'étape suivante. Où n(d) est la décision de sortie de cette étape particulière donnant sa prédiction de nombres/classes continus en cas de régression/classification. n(a) qui va servir d'entrée au prochain 'Attentative Transformer' où le cycle suivant commence.

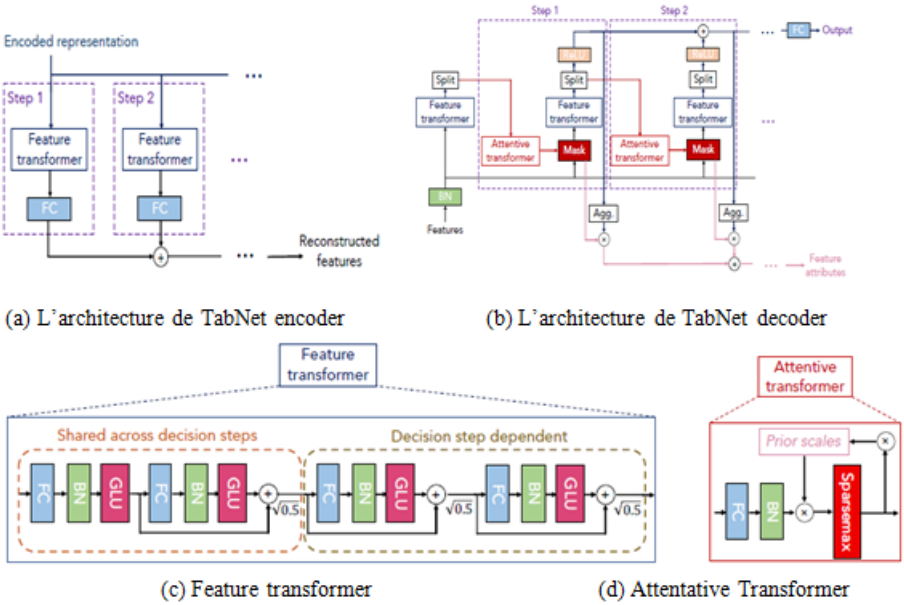


FIGURE 3.15. Architecture TabNet

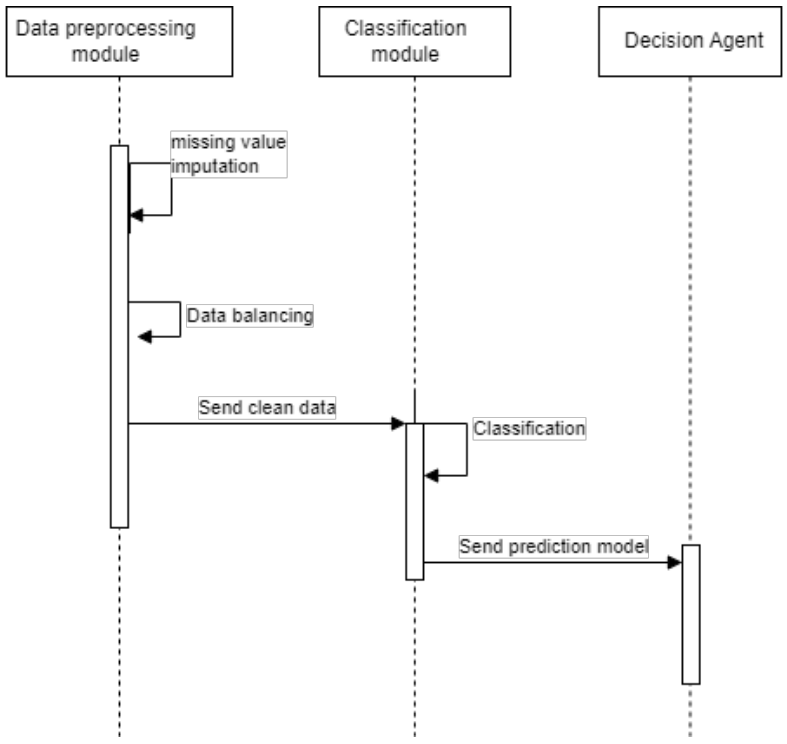


FIGURE 3.16. Diagramme de séquence de la phase "génération du modèle de prédiction"

3.4.4.2 Acquisition des données

Cette phase permet d'acquérir les données des patients et d'assurer leur anonymisation et les sauvegarder grâce à trois agents (interface agent, data collection agent et data protection agent).

1. **Agent Interface (Interface Agent : IA)**

   Cet agent permet d'assurer la communication entre le système et l'utilisateur (patients et médecins). Son rôle principal est l'acquisition et la transmission des données (figure 3.17). Il est composé de :

   — Un module de communication (agent-utilisateur) qui assure la communication et le transfert de données entre les utilisateurs et le système.

   — Un module de communication (agent-agent) permettant de transférer les données et les résultats avec l'agent DA.

   — Un module de traitement pour filtrer les données acquises (données erronées).

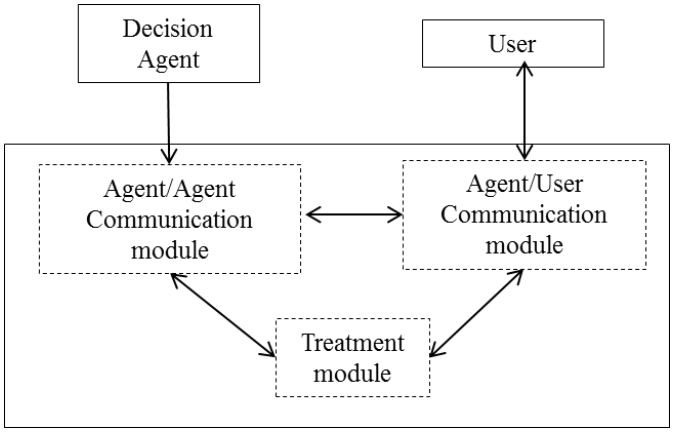


FIGURE 3.17. Architecture de IA

2. **Agent collection de données (Data Collection Agent : DCA)**

   Le rôle principal de cet agent est de collecter les données médicales auprès de l'agent d'interface et assurer leur vérification (figure 3.18). Il est composé de :

   — Un module de communication qui assure les échanges entre les différents agents du système.

   — Un module d'acquisition de données pour la collecte et la vérification de données.

   — Un ensemble de données médicales pour le stockage des données acquises.

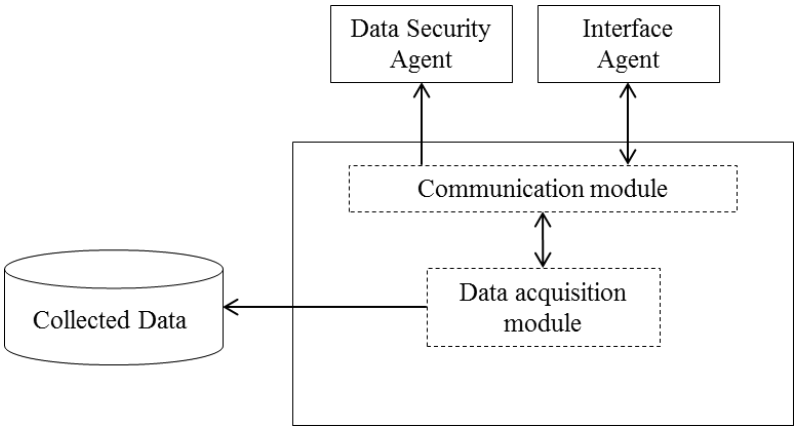


FIGURE 3.18. Architecture de DCA

3. **Agent de protection des données (Data Protection Agent : DPA)**

   L'anonymisation et la sécurité des données consistent à protéger les données médicales des patients contre la lecture et l'utilisation par des personnes non autorisées. C'est une tâche importante en médecine. Pour l'assurer, nous proposons d'utiliser l'agent (data privacy) qui se compose de (figure 3.19) :

   — Un module de communication.

   — Un module de collecte de données.

   — Un module d'anonymisation des données pour assurer la confidentialité des données.

   — Un module de protection pour la sécurité des données.

   — Un ensemble de données pour le stockage de données anonymes.

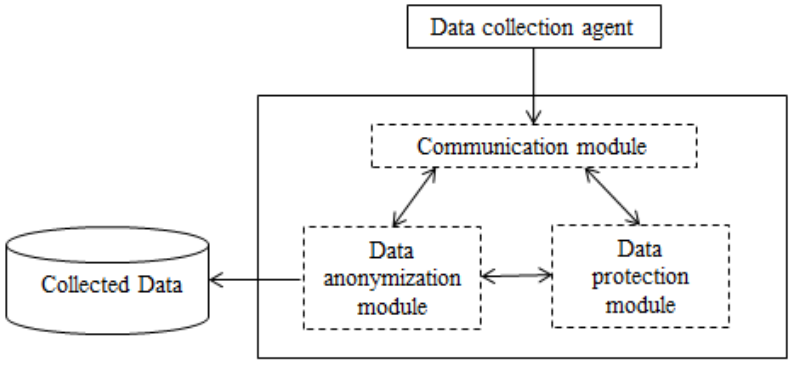


FIGURE 3.19. Architecture de DPA

3.4.4.3 Evaluation et prise de décision

Dans cette phase l'agent (**Decision Agent : DA**) est chargé d'assurer la prise de décision et l'évaluation du modèle de prédiction. C'est un agent cognitif, il reçoit les données prétraitées, applique le modèle de prédiction sur ces dernières et assure l'évaluation et la validation des résultats (figures 3.20 et 3.21). Il est composé de :

— Module de communication.

— Module d'évaluation/validation.

— Module de décision.

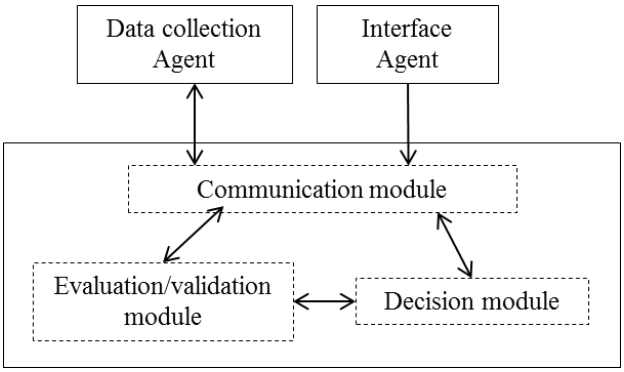


FIGURE 3.20. Architecture de DA

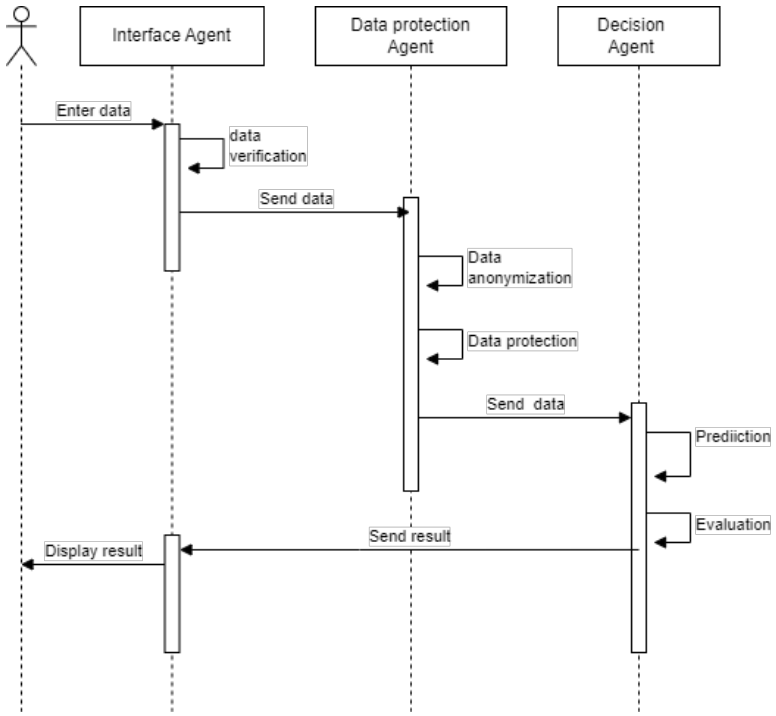


FIGURE 3.21. Diagramme de séquence des phases "acquisition des donées, evaluation et prise de décision"

3.4.5 Implémentation et résultats

Dans cette section, nous présentons la mise en œuvre et le processus du système proposé, y compris les outils de développement et les résultats expérimentaux.

3.4.5.1 Outils et plateformes utilisés

Dans ce travail, les résultats sont obtenus grâce aux environnements Python présentés ci-dessous.

1. **Sckit-learn**

   Scikit-learn ou sklearn est une bibliothèque open source pour le ML en Python. Elle prend en charge l'apprentissage supervisé et non supervisé. Elle fournit également divers outils pour le prétraitement des données, la sélection, la modification, l'optimisation et l'évaluation d'un modèle, et de nombreux autres utilitaires. Ses principales fonctionnalités incluent :

   (a) Méthodes de prise de décision algorithmique.

   (b) Algorithmes prenant en charge l'analyse prédictive allant de la simple régression linéaire à la reconnaissance de modèles de réseaux de neurones.

   (c) Interopérabilité avec les bibliothèques NumPy, pandas et matplotlib.

2. **PyTorch**

   PyTorch est un framework Python basé sur la librairie Torch. Il fournit deux fonctionnalités de haut niveau :

   (a) Calcul du tenseur (comme NumPy) avec une forte accélération GPU.

   (b) Réseaux de neurones profonds construits sur un système d'autogradation sur bande.

3. **Optuna**

   Optuna est un framework logiciel d'optimisation automatique d'hyperparamètres. Il comporte une API utilisateur impérative de style défini par exécution. Grâce à notre API de définition par exécution, le code écrit avec Optuna bénéficie d'une grande modularité, et l'utilisateur d'Optuna peut construire dynamiquement les espaces de recherche des hyperparamètres. Il est facile à utiliser et prend en charge une variété d'algorithmes d'optimisation.

4. **Python Agent DEvelopment (PADE) framework**

   PADE est un framework open-source pour le développement, l'exécution et la gestion d'environnements de systèmes multi-agents de calcul distribué. PADE est écrit à 100% en langage Python et utilise les bibliothèques Twisted pour implémenter la communication entre les nœuds du réseau. PADE offre plusieurs fonctionnalités :

   (a) Abstraction des agents et leurs comportements en utilisant des concepts d'orientation objet.

   (b) Interaction avec les bases de données.

   (c) Manipulation des échanges de messages entre les agents en utilisant le protocole FIPA-ACL.

   (d) Filtrage des messages multi-agents.

3.4.5.2 Résultats expérimentaux et discussion

Pour évaluer l'efficacité de l'optimisation basée sur l'algorithme NSGA-II du modèle de classification TabNet, plusieurs comparaisons ont été menées. Les résultats obtenus seront discutés ci-dessous.

1. **Comparaison des résultats obtenus avec et sans application de NSGA-II**

   Le tableau 3.11 résume les résultats obtenus en termes de PPV, score-F1, AUC et exactitude de valeurs respectives de : 73%, 71%, 72%, 73% et 92%. Ces résultats montrent que l'application de l'algorithme d'optimisation NSGA-II améliore les performances du modèle de classification TabNet. Cependant, le modèle TabNet obtient un rappel, qui est également une mesure de performance très importante, légèrement plus élevé (72 %) que celui de l'approche proposée (71%).

TABLE 3.11. Résultats obtenus avec et sans application de NSGA-II

Modèle	PPV	Rappel	score-f1	AUC	Exactitude
TabNet	69%	72%	71%	65%	90%
NSGAII-TabNet	73%	71%	72%	73%	92%

Le tableau 3.13 présente les hyperparamètres du modèle de classification TabNet sélectionnés en appliquant l'algorithme d'optimisation NSGA-II, y compris la fonction de masquage, l'architecture, la fonction d'optimisation, le nombre d'époques ... etc. A partir des résultats du tableau 3.11 et les paramètres présentés dans le tableau 3.13, nous pouvons remarquer que la modification des hyperparametres du modèle conduit à la modification des mesures de performance.

2. **Comparaison des résultats obtenus avec d'autres approches de la littérature**

   Pour mieux évaluer l'approche proposée, une deuxième comparaison a été menée. Elle vise à comparer TabNet optimisé par NSGA-II et d'autres modèles de la littérature. Les expérimentations ont été menées dans les mêmes conditions (dataset, sous-ensemble d'apprentissage/test et même objectif). En observant le tableau 3.12 et la figure 3.22, nous pouvons remarquer que la pluparts des métriques montrent une amélioration de la performance de la prédiction. Dans l'ensemble, la combinaison du modèle de classification TabNet et la méthode d'optimisation NSGA-II s'avère extrêmement compétitive et surpasse les autres méthodes de classification. Cependant, les méthodes DT, CBR et la logique floue ont surpassé l'approche proposée en termes de score-f1 et les méthodes k-SVM et RF la surpassent en termes de rappel.

TABLE 3.12. Comparaison de l'approche proposée avec d'autres approches

Modèle	PPV	Rappel	score-f1	AUC	Exactitude
KNN	61%	68%	63%	65%	84%
k-SVM	68%	74%	71%	72%	89%
RF	72%	73%	72%	52%	91%
DT [118]	-	-	95.06%	-	90.63%
CBR [118]	-	-	92.54%	-	86.26%
Logique floue [118]	-	-	95.64%	-	91.64%
NSGAII-Tabnet	73%	71%	72%	73%	92%

TABLE 3.13. Hyperparamètres sélectionnés par NSGA-II

Paramètre	Description	Valeur
'mask_type'	il s'agit de la fonction de masquage à utiliser pour sélectionner les entités.	entmax
'n_steps'	Nombre d'étapes dans l'architecture.	2
gamma	coefficient de réutilisation des fonctionnalités dans les masques. Les valeurs vont de 1,0 à 2,0.	1.0
'n_shared'	Nombre d'unités linéaires fermées partagées à chaque étape	3
Optimizer	Fonction d'optimisation de Pytorch.	Adam
'lambda_sparse'	Il s'agit du coefficient de perte de parcimonie supplémentaire.	0.00028488042227549603
patienceScheduler	Pour un arrêt précoce quand "patience" s'arrête de s'améliorer.	7
patience	Nombre d'époques consécutives sans amélioration avant d'effectuer un arrêt précoce.	30
epochs	Nombre d'époques d'entrainement.	70

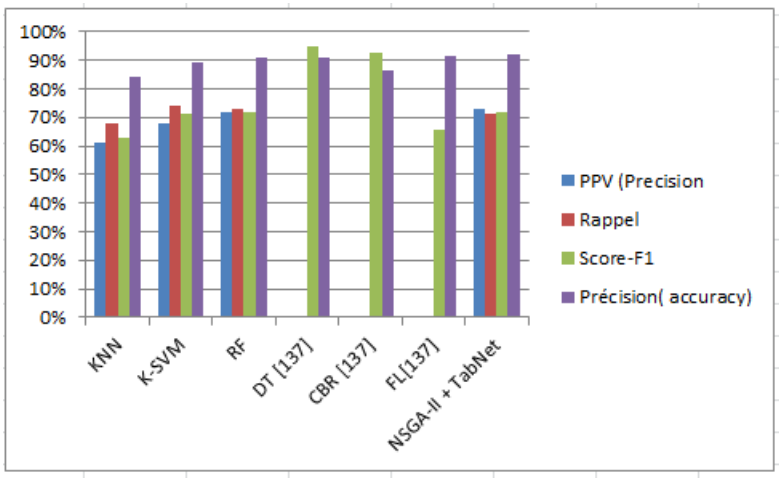


FIGURE 3.22. Comparaison de l'approche proposée avec d'autres approches

3. **La courbe ROC**

Pour valider les performances de classification de la combinaison du modèle TabNet avec la méthode d'optimisation NSGA-II, nous avons utilisé également la courbe ROC. La figure 3.23 montre que l'approche proposée surpasse la plupart des méthodes étudiées en termes de AUC avec une valeur de 73%, qui est considérée comme acceptable pour la prédiction des patients atteints du COVID-19 ayant besoin d'accéder aux soins intensifs [142].

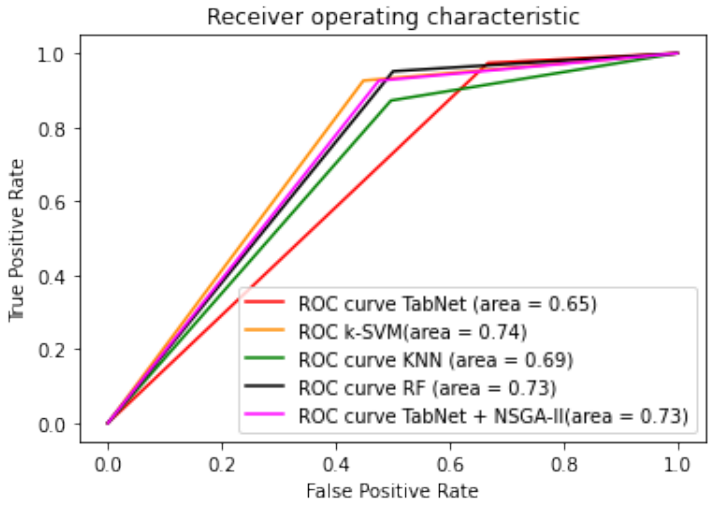


FIGURE 3.23. La courbe ROC des différents modèles

4. **Caractéristiques de l'approche proposée avec et sans SMA**

   L'approche SMA ajoute de nouvelles fonctionnalités (tableau 3.14) qui permettent au système proposé d'atteidre ses objectifs de façon autonome.

TABLE 3.14. Caractéristiques de l'approche proposée avec et sans SMA

	Autonomie	Communication	Interactions	Traitement parallèle	Raisonnement
Sans SMA	×	×	×	×	✓
Avec SMA	✓	✓	✓	✓	✓

Les figures 3.24, 3.25, 3.26, 3.27 montrent le fonctionnement de notre SMA. Après le lancement du SMA, les agents sont créés. Ils communiquent entre eux par échange de messages, le modèle de prédition TabNet optimisé par l'algorithme NSGA-II est appliqué par l'agent prise de décision (DA) sur les données de l'utilisateur recueillis par l'agent interface (IA). Ce dernier affiche le résultat de la prédiction à l'utilisateur.

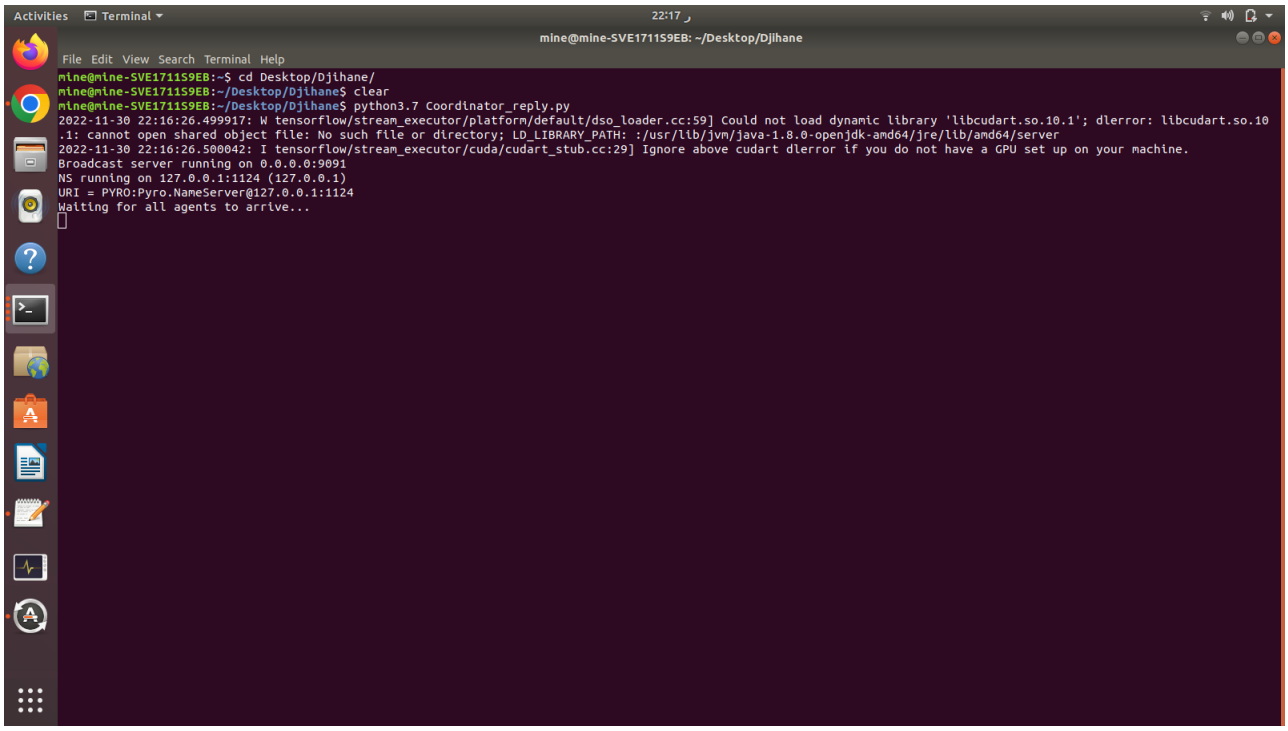


FIGURE 3.24. Lancement du SMA

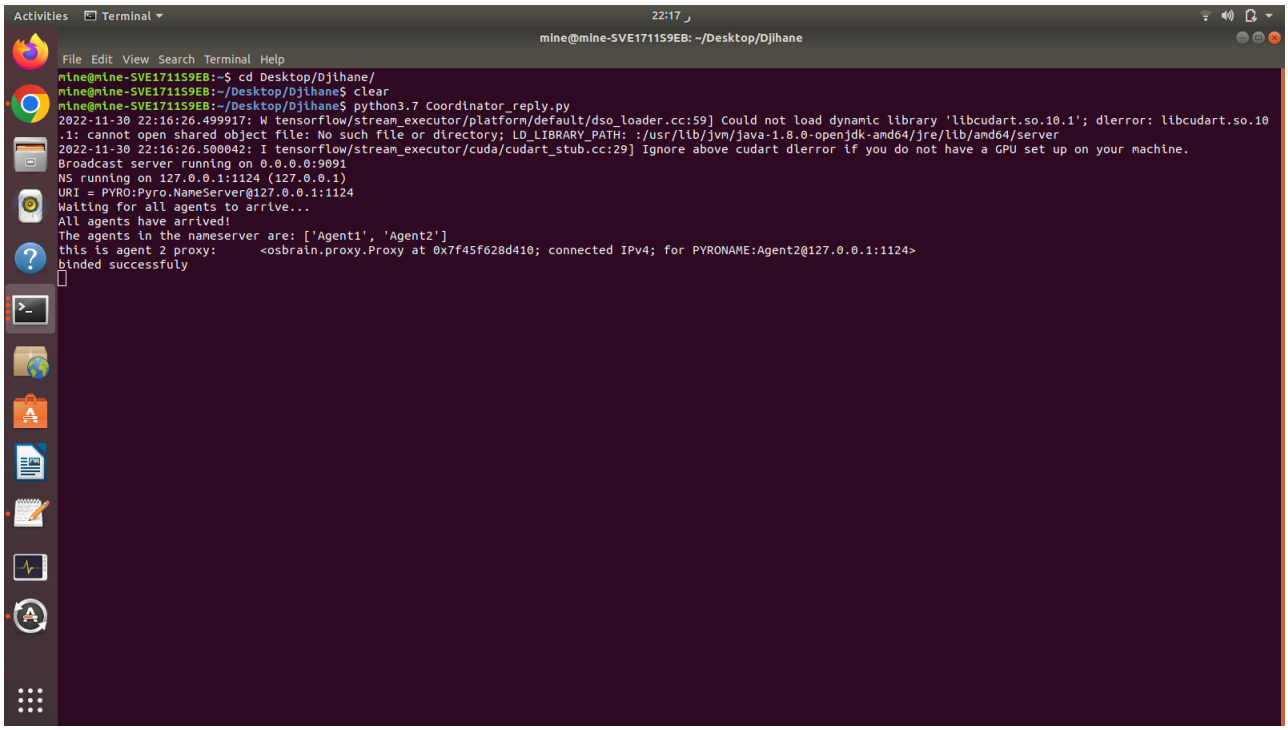


FIGURE 3.25. Création des agents

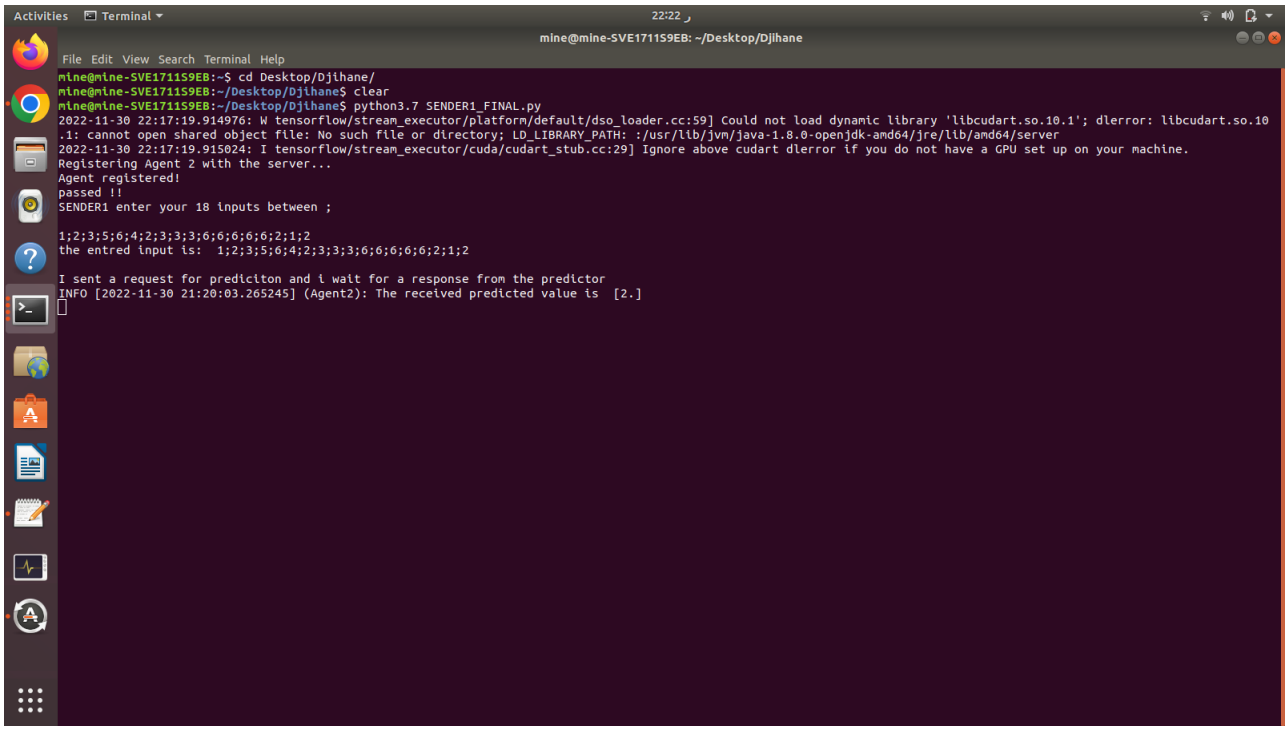


FIGURE 3.26. Communication entre les agents

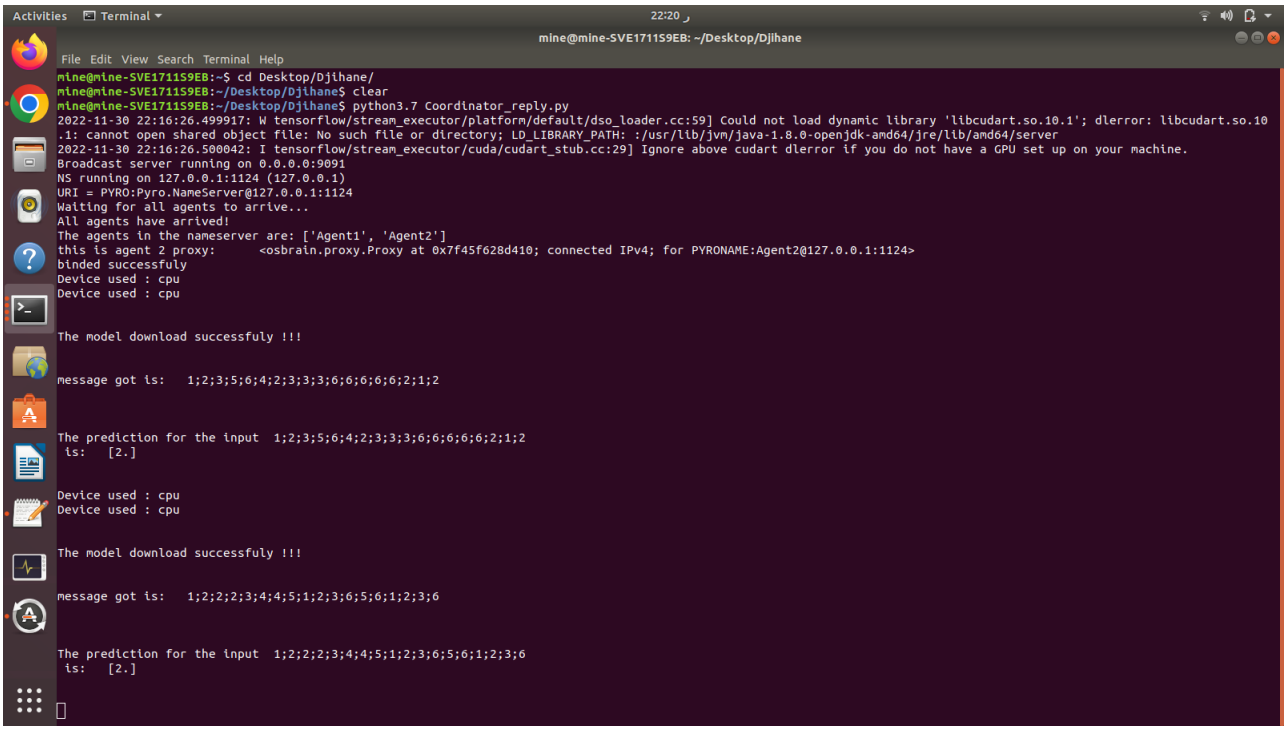


FIGURE 3.27. Sauvergarde du modèle de prédiction, traitement et affichage du résultat

3.5 Conclusion

L'IA est largement utilisée dans le domaine biomédical et sanitaire et elle ne cesse de les révolutionner. L'un des plus grands défis de l'application de l'IA dans ces domaines est de créer des systèmes d'aide à la décision et des systèmes de prédiction précis et efficaces en termes de temps de calcul et de performance. Au cours de la dernière décennie, de nombreux travaux de recherche ont été menés dans le domaine médical pour cette raison. L'application des techniques de l'IA telles que ML, DL, SMA et métaheuristiques ont montré une remarquable capacité à améliorer l'exactitude de la prédiction médicale.

Dans ce chapitre nous avons présenté nos principales contributions appliquant les techniques de l'IA pour le diagnostic et la prédiction médicale. Nous avons conduit une étude comparative entre les méthodes d'apprentissage automatique les plus utilisées dans la littérature. En outre, nous avons proposé une approche basée sur la méthode de réduction de fonctionnalités PCA et la régression logistique pour la classification des tumeurs mammaires. Pour réaliser ce système deux bases de données publiques (WDBC et WOBC) du référentiel d'apprentissage automatique UCI ont

été utilisées. Les résultats expérimentaux ont montré la performance de cette approche en termes d'exactitude, sensibilité, f1-score et PPV par rapport aux autres travaux basés sur ML et DL utilisant les même BD. Cependant, l'application de cette approche sur d'autres bases de données plus volumineuses n'a pas prouvé son efficacité. La deuxième contribution est un système de prédiction de la COVID-19 basé sur les SMA et le modèle de DL TabNet optimisé par l'algorithme NSGA-II. Les résultats obtenus ont montré que cette approche a surperformé les autres modèles de la littérature.

Conclusion et perspectives

La prédiction médicale est un défi important pour les cliniciens, car elle a une influence directe sur leur pratique quotidienne. Au cours de la dernière décennie, le taux de mortalité a considérablement augmenté, ce qui a nécessité des méthodes et des outils pour une détection précise et précoce des pathologies.

Nos travaux visent à montrer l'apport de la technologie au bénéfice des professionnels de santé en tant qu'outil d'analyse et de calcul, mais surtout un outil d'aide au diagnostic dans plusieurs disciplines médicales. Nous avons proposés des solutions innovantes qui visent à simplifier le traitement des données collectées, et des algorithmes capable de traduire celles-ci en informations claires et exploitables par les professionnels de la santé.

Cette thèse intègre trois principaux concepts de l'IA, à savoir les SMA, le ML et les métaheuristiques dans la prédiction médicale. Les travaux présentés, mettent l'accent sur les systèmes de diagnostic assisté par ordinateur pour assister les médecins à poser un diagnostic précis. Les problèmes traités dans cette thèse ont donné naissance à deux principales contributions et une étude comparative qui peuvent être résumées comme suit :

- L'un des défis les plus importants du ML est de développer des méthodes de classification précises et efficaces en termes de calcul. Nous avons utilisé MLP, L- SVM, K-SVM, DT, RF, KNN, LR et NB sur l'ensemble de données WDBC et comparé les performances de ces algorithmes en termes d'efficacité, pour déterminer la meilleure méthode de ML utilisée pour la classification binaire des tumeurs mammaire. A travers cette étude expérimentale, nous avons atteint la valeur la plus élevée de précision (98 %) avec MLP et LR. Ce qui prouve leur efficacité pour la classification des tumeurs mammaires.
- Notre première contribution consiste à développer un CAD basé sur PCA pour l'extraction des caractéristiques et la régression logistiques pour la classification des tumeurs mam-

maires. Les résultats expérimentaux sur les bases de données WDBC et WOBC démontrent, que l'extraction de caractéristiques en appliquant la méthode PCA est avantageuse, car elle optimise les performances de classification par LR en améliorant la qualité des données, en diminuant le nombre de caractéristiques sans perdre les principales informations objectives de l'ensemble de données d'origine. En conséquence, il a l'avantage de réduire le coût de calcul et le temps de traitement. Le système proposé a surpassé les autres travaux de recherche proposés dans la littérature en atteignant une exactitude de 1 et 0.97 pour WDBC et WOBC respectivement. Cependant, l'application de cette approche sur d'autres bases de données plus volumineuses, n'a pas prouvé son efficacité (problèmes d'overfitting et underfitting).

- En se basant sur les limites de la première contribution, nous avons opté pour le paradigme du système multi-agents, pour tirer profit de ses caractéristiques (autonomie, communication, traitement parallèle et collaboration), DL et les méta-heuristiques, afin de résoudre les défis de l'application des techniques de l'IA pour la gestion de la pandémie Covid-19.

Nous l'avons vu tout au long de nos études et travaux, l'IA, à travers l'ensemble de ses composantes est un vecteur de progrès, notamment dans le domaine de la médecine. A travers la littérature scientifique, nous avons pu identifier un grand nombre de solutions ayant parfois été mises en application de façon concrète en médecine ; le nombre d'articles publiés peut en témoigner. Nous avons également constaté que l'informatique prédictive fait face à la méfiance des médecins. Ces derniers ne voient pas toujours d'un bon œil l'idée qu'une machine puisse remplacer, au moins partiellement, leur expertise. Certaines interrogations méritent d'être posées.

— *Comment convaincre un professionnel que la machine est un assistant virtuel et non un adversaire exterminateur ?*

— *Comment ajouter de l'éthique à un raisonnement trivial d'une machine qui peut être erroné sans décrédibiliser le praticien dans ses fonctions qu'il sait pourtant faire sans cet outil ?*

— *Comment garantir le bon fonctionnement de la machine en tenant compte de la qualité des données en termes de sécurité, confidentialité et interopérabilité en conformité avec la loi en vigueur ?*

Toutes ces questions doivent faire l'objet de travaux enrichis car la confiance du monde médicale passera par là. Pour répondre à ces questions, plusieurs améliorations pourraient être apportées

aux travaux effectués dans cette thèse et des axes de recherche nécessitant des investigations plus approfondies. Nous considérons les extensions suivantes :

— Nous visons à appliquer nos travaux dans un système de diagnostic clinique réel du cancer du sein, pour assister les médecins dans le processus de la prise de décision et à les appliquer au diagnostic d'autres maladies.

— La sécurité et la confidentialité des données des patients fait partie des défis auxquels l'IA est confrontée. Nous proposons d'aborder ce thème en développant l'"Agent protection des données (DPA)", pour assurer la protection des données et améliorer la qualité de notre IDSS.

— Orienter l'architecture globale vers une vision IoT, pour améliorer l'accès au diagnostic médical.

— Améliorer la performance de la deuxième contribution et l'évaluer sur de nouvelles datasets et la comparer à d'autres travaux proposés aux challenges mondiaux. Cela permettra de mieux vulgariser l'application de l'IA dans le diagnostic et le pronostic médical.

Annexe A

Description des datasets

A.1 WDBC

La Wisconsin Diagnosis Breast Cancer dataset (WDBC) est extraite du référentiel d'apprentissage automatique de l'Université de Californie à Irvine (UCI). Elle comporte 569 cas dont 357 (62,7%) bénins et 212 (37,3%) malins. Elle comporte 32 attributs, obtenues à partir d'aspirations à l'aiguille fine (FNA), présentés dans le tableau A.1.

TABLE A.1. Description de la dataset WDBC

No	Attribut	Description
1	Radius	moyenne des distances du centre aux points du périmètre
2	Texture	écart type des valeurs d'échelle de gris
3	Perimeter	
4	Area	
5	Smoothness	variation locale des longueurs de rayon
6	Compactness	perimeter^2 / area - 1.0
7	Concavity	sévérité des parties concaves du contour
8	Concave points	nombre de parties concaves du contour
9	Symmetry	
10	Fractal dimension :	coastline approximation - 1

A.2 WOBC

L'ensemble de données WOBC contient 699 enregistrements et neuf caractéristiques obtenues à partir d'aspirations à l'aiguille fine (FNA). Le tableau A.2 présente sa description.

TABLE A.2. Description de la dataset WOBC

No	Attribut	valeur
1	Clump Thickness	1 - 10
2	Uniformity of Cell Size standard deviation of gray-scale values	1 - 10
3	Uniformity of Cell Shape	1 - 10
4	Marginal Adhesion	1 - 10
5	Single Epithelial Cell Size	1 - 10
6	Bare Nuclei	1 - 10
7	Bland Chromatin	1 - 10
8	Normal Nucleoli	1 - 10
9	Mitoses	1 - 10
10	Class	2 for benign, 4 for malignant

A.3 COVID-19 patient pre-condition dataset

L'ensemble de données "COVID-19 patient pre-condition dataset" de Kaggle comprend 566602 enregistrements et 23 attributs. Sa descriscription est résumée comme suit.

TABLE A.3. Description de COVID-19 patient pre-condition dataset

No	Attribut	Description
1	Patient id	Chaîne de caractères
2	Sex	Femelle 1, Male 2
3	patient type	Patient non hospitalisé 1, patient hospitalisé 2
4	entry_date	Date d'hospitalisation
5	date_symptoms	Date de l'apparition des symptômes
6	date_died	Date manquante = 9999-99-99
7	Intubed	Oui 1, Non 2, Donnée manquante ou NA 97, 98,99
8	Pneumonia	Oui 1, Non 2, Donnée manquante ou NA 97,98,99
9	Age	
10	Pregnancy	Oui 1, Non 2, Donnée manquante ou NA 97, 98,99
11	Diabetes	Oui 1, Non 2, Donnée manquante ou NA 97, 98,99
12	Copd	Bronchopneumopathie chronique obstructive (Oui 1, Non 2, Donnée manquante ou NA 97, 98,99)
13	Asthma	Oui 1, Non 2, Donnée manquante ou NA 97,98,99
14	inmsupr	Identifie si le patient souffre d'immunosuppression (Oui 1, Non 2, Donnée manquante ou NA 97, 98,99)
15	hypertension	Oui 1, Non 2, Donnée manquante ou NA 97,98, 99
16	other diseases present	Oui 1, Non 2, Donnée manquante ou r NA 97,98, 99
17	cardiovascular	Oui 1, Non 2, Donnée manquante ou NA 97,98, 99
18	obesity	Oui 1, Non 2, Donnée manquante ou NA 97,98, 99
19	renal_chronic	Oui 1, Non 2, Donnée manquante ou NA 97,98, 99
20	tobacco	Oui 1, Non 2, Donnée manquante ou NA 97,98, 99
21	contact with other covid	Oui 1, Non 2, Donnée manquante ou NA 97,98, 99
22	covid result	Positif 1, Négatif 2, en attente 3
23	Icu	Identifie si le patient doit entrer dans une unité de soins intensifs (Oui 1, Non 2, Donnée manquante ou NA 97, 98, 99)

Annexe B

Liste des publications

B.1 Revues Internationales

Djihane Houfani, Sihem Slatnia, Okba Kazar, Noureddine Zerhouni, Ikram Remadna, and Hamza Saouli, ***"Breast cancer diagnosis using machine learning techniques : a comparative study"***, *Medical Technologies Journal,* Volume : 4, Issue : 2, April-June 2020, pp : 535-544 DOI : http ://medtech.ichsmt.org/index.php/MTJ/article/view/248

Djihane Houfani, Sihem Slatnia, Okba Kazar, Hamza Saouli, and Abdelhak Merizig, ***"Artificial intelligence in healthcare : a review on predicting clinical needs"***, *International Journal of Healthcare Management,* volume 15(3), pp. 267-275, 2022. DOI : 10.1080/20479700.2021.1886478

Djihane Houfani, Sihem Slatnia, Okba Kazar, Ikram Remadna, Hamza Saouli,Guadalupe Ortiz, and Abdelhak Merizig, ***"An Improved Model for Breast Cancer Diagnosis by Combining PCA and Logistic Regression Techniques"***, *International Journal of Computing and Digital Systems*, 2023, https ://journal.uob.edu.bh/handle/123456789/4760.

B.2 Conférences Internationales

Djihane Houfani, Sihem Slatnia, Okba Kazar, Noureddine Zerhouni, Abdelhak Merizig, and Hamza Saouli, ***"Machine Learning Techniques for Breast Cancer Diagnosis : Literature Review"***, The International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2019), Marrakech-Morocco, 08-11 July 2019, pp. 247–254. DOI : 0.1007/978-3-

030-36664-3_28.

Djihane Houfani, Sihem Slatnia, Okba Kazar, Noureddine Zerhouni, Ikram Remadna, and Hamza Saouli, ***"A practical comparative study of machine learning algorithms for breast cancer diagnosis",*** International congress on Health Sciences and Medical Technologies, Tlemcen, Algeria 5-7 December 2019 https ://doi.org/10.26415/978-9931-9446-2-1.

Salah Eddine Henouda, Fatima Zohra Laallam, Okba Kazar, Saad Harous and Djihane Houfani, ***"On the effectiveness of Dimensionality Reduction Techniques on High Dimensionality Datasets"***, 12th International Conference on Information Systems and Advanced Technologies "ICISAT'2022".

Djihane Houfani, Sihem Slatnia, Okba Kazar, Hamza Saouli , Salah Eddine Henouda ,Ikram Remadna, and Meftah Zouai, ***"TabNet Based Prediction Model for ICU admission in Covid-19 patients"***, The International Symposium on iNnovative Informatics, Biskra, Algeria, December 7-8th, 2022.

Bibliographie

[1] Louis Frécon and Okba Kazar. *Manuel d'intelligence artificielle*. PPUR Presses polytechniques, 2009.

[2] Sachin S Kamble, Angappa Gunasekaran, Milind Goswami, and Jaswant Manda. A systematic perspective on the applications of big data analytics in healthcare management. *International Journal of Healthcare Management, Taylor & Francis*, 2018.

[3] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and SS Iyengar. Computational health informatics in the big data age : a survey. *ACM Computing Surveys (CSUR), ACM New York, NY, USA*, 49(1) :1–36, 2016.

[4] J. Ferber. *Les systèmes Multi-Agents, vers une intelligence collective*. InterEditions, 1995.

[5] Stuart Russell and Norvig Peter. *Intelligence artificielle*. Pearson Education France, 2010.

[6] Michael Wooldridge and Nicholas R. Jennings. Intelligent agents : Theory and practice. *The knowledge engineering review*, 10 :115–152, 1995.

[7] Imed Jarras and Brahim Chaib-Draa. Aperçu sur les systèmes multi-agents. rapport technique. *Montréal : CIRANO (Centre Interuniversitaire de Recherche et Analyse des Organisations)*, juillet 2002.

[8] Jim Doran, Stan Franklin, Nick Jennings, and Timothy Norman. On cooperation in multi-agent systems. *The Knowledge Engineering Review, Cambridge University Press New York, USA*, 12 :309–314, 1997.

[9] Lazhar Benoudina. Modélisation et simulation basées multi-agents du contrôle de prossesus industriel. *Mémoire de magister : Informatique. Université de Skikda*, 2009.

[10] Sandip Sen and Weiss Gerhard. Learning in multiagent systems. *A modern approach to distributed artificial intelligence*, pages 259–298, 1999.

[11] Junling Hu and P. Wellman Michael. Multi-agent reinforcement learning : theoretical framework and an algorithm. *ICML*, 98 :242–250, 1998.

[12] Amit Kumar and Bikash Kanti Sarkar. A case study on machine learning and classification. *International Journal of Information and Decision Sciences*, 9 :179–208, 1997.

[13] François Chollet. *Deep Learning with Python*. Manning Publications, 2018.

[14] Géron Aurélien. *Hands-On Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017.

[15] B. Boehmke and B. Greenwell. *Hands-On Machine Learning with R*. the R series, 2020.

[16] Ian T. Jolliffe and Cadima Jorge. Principal component analysis : a review and recent developments. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 374 :20150202, 2016.

[17] Sola Jorge and Sevilla Joaquin. Importance of input data normalization for the application of neural networks to complex industrial problems. *International Journal of Computer Information Systems and Industrial Management Applications*, 6 :257–269, 2014.

[18] S. C. Nayak, B. Misra Bijan, and Himansu Sekhar Behera. Impact of data normalization on stock index forecasting. *IEEE Transactions on nuclear science*, 44 :1464–1468, 1997.

[19] Samina Khalid, Khalil Tehmina, and Nasreen Shamila. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014.

[20] Vipin Kumar and Minz Sonajharia. Feature selection : a literature review. *SmartCR*, 3 :211–229, 2014.

[21] Manoranjan Dash and Liu Huan. Feature selection for classification. *Intelligent Data Analysis, Elsevier*, 1 :131–156, 1997.

[22] Siti Hawa Apandi, Sallim Jamaludin, and Mohamed Rozlina. A survey on technique for solving web page classification problem. In *IOP Conference Series : Materials Science and Engineering*, page 012036. IOP Publishing, 2020.

[23] Jesse Davis and Goadrich Mark. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[24] Thomas Davenport and Kalakota Ravi. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6 :94, 2019.

[25] Kenneth Sörensen and Glover Fred. Quantifying complexity theory. *Encyclopedia of operations research and management science*, 62 :960–970, 2013.

[26] Metropolis Nicholas, W. Rosenbluth Arianna, N. Rosenbluth Marshall, and H. Teller Augusta. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21 :1087–1090, 1953.

[27] Scott Kirkpatrick, Gelatt Jr C. Daniel, and P. Vecchi Mario. Optimization by simulated annealing. *Science*, 220 :671–680, 1983.

[28] Černý Vladimir. Thermodynamical approach to the traveling salesman problem : an efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45 :41–51, 1985.

[29] Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13 :533–549, 1986.

[30] Thomas A. Feo and GC Resende. Mauricio. A probabilistic heuristic for a computationally difficult set covering problem. *Operations Research Letters*, 8 :67–71, 1989.

[31] Thomas A. Feo and GC Resende Mauricio. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6 :109–133, 1995.

[32] Patrick Siarry. *Métaheuristiques*. Editions Eyrolles, 2014.

[33] John H. Holland. *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[34] Kenneth. De Jong. Learning with genetic algorithms : An overview. *Machine learning*, 3 :121–138, 1988.

[35] Yann COLLETTE and Patrick SIARRY. *Optimisation multiobjectif*. Editions Eyrolles, 2002.

[36] Zong Woo Geem, Joong Hoon Kim, and Gobichettipalayam Vasudevan Loganathan. A new heuristic optimization algorithm : Harmony search. *Simulation*, 76 :60–68, 2001.

[37] Askarzadeh Alireza and Esmat Rashedi. Harmony search algorithm : Basic concepts and engineering applications. In *Intelligent Systems : Concepts, Methodologies, Tools, and Applications*, pages 1–30. IGI Global, 2018.

[38] Leroy HOOD and Mauricio FLORES. A personal view on systems medicine and the emergence of proactive p4 medicine : predictive, preventive, personalized and participatory. *New biotechnology*, 29 :613–624, 2012.

[39] Mauricio Flores, Gustavo Glusman, Nathan D Price, Kristin Brogaard, and Leroy Hood. P4 medicine : how systems medicine will transform the healthcare sector and society. *Personalized medicine*, 10 :565–576, 2012.

[40] Leroy HOOD, Rudi BALLING, and Charles AUFFRAY. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnology journal*, 7 :992–1001, 2012.

[41] Jacques RUFFIÉ. *Naissance de la médecine prédictive*. Odile Jacob, 1993.

[42] Martine BATT, Alain TROGNON, and Denis VERNANT. Quand l'argument effleure la conviction : Analyse interlocutoire dune croyance dans un entretien de médecine prédictive. *Psychologie de linteraction*, pages 17–18, 2004.

[43] Thierry Mathieu, Laurent Bermont, Jean-Christophe Boyer, Céline Versuyft, Alexandre Evrard, Isabelle Cuvelier, Remy Couderc, and Katell Peoc'h. Champs lexicaux de la médecine prédictive et personnalisée. *Annales de Biologie Clinique*, 70 :651–658, 2012.

[44] J. F. Jekel, D. L. Katz, J. G. Elmore, and D. Wild. *Epidemiology, biostatistics and preventive medicine*. Elsevier Health Sciences, 2007.

[45] Isaac S. CHAN and S. GINSBURG, Geoffrey. Personalized medicine : progress and promise. *Annual review of genomics and human genetics*, 12 :217–244, 2011.

[46] Alain CLAEYS and Jean-Sébastien VIALATTE. Les progrès de la génétique : vers une médecine de précision ? les enjeux scientifiques, technologiques, sociaux et éthiques de la médecine personnalisée. *Office parlementaire d'évaluation des choix scientifiques et technologiques*, 22, 2014.

[47] Hood Leroy, A. Flores Mauricio, R. Brogaard Kristin, and D. Price Nathan. Systems medicine and the emergence of proactive p4 medicine : predictive, preventive, personalized and participatory. *Handbook of Systems Biology,Elsevier Inc*, 22 :445–467, 2013.

[48] Dumez Hervé, Minvielle Etienne, and Marrauld Laurie. La médecine participative note complémentaire n° 2 du rapport " État des lieux de l'innovation en santé numérique. novembre 2015.

[49] Vepa Atamuradov, Kamal Medjaher, Pierre Dersin, Benjamin Lamoureux, and Noureddine Zerhouni. Prognostics and health management for maintenance practitioners-review, implementation and tools evaluation. *International Journal of Prognostics and Health Management*, 8(3) :1–31, 2017.

[50] Arpit Bhardwaj and Aruna Tiwari. Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications, Elsevier*, 42 :4611–4620, 2015.

[51] Ashraf Osman Ibrahim and Siti Mariyam Shamsuddin. Intelligent breast cancer diagnosis based on enhanced pareto optimal and multilayer perceptron neural network. *International Journal of Computer Aided Engineering and Technology, Inderscience*, 10 :543–556, 2018.

[52] Na Liu, Er-Shi Qi, Man Xu, Bo Gao, and Gui-Qiu Liu. A novel intelligent classification model for breast cancer diagnosis. *Information Processing and Management, Elsevier*, 56 :609–623, 2019.

[53] Nawel Zemmal, Nabiha Azizi, Nilanjan Dey, and Mokhtar Sellami. Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification. *Journal of Medical Imaging and Health Informatics*, 6 :53–62, 2016.

[54] Abdulkader Helwan, John Bush Idoko, and Rahib H. Abiyev. Machine learning techniques for classification of breast tissue. In *9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW*, pages 402–410. Elsevier, 2017.

[55] Haifeng Wang, Bichen Zheng, Sang Won Yoon, and Hoo Sang Ko. Ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research, Elsevier*, 267 :687–699, 2017.

[56] Kemal Polat and Umit Sentürk. A novel ml approach to prediction of breast cancer : Combining of mad normalization, kmc based feature weighting and adaboostm1 classifier. In *International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–4. IEEE, 2018.

[57] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12 :e0177544, 2017.

[58] Fabio Alexandre Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)*, pages 2560–2567. IEEE, 2016.

[59] Ahmed M. Abdel-Zaher and Ayman M. Eldeib. Breast cancer classification using deep belief networks. *Expert Systems with Applications, Elsevier*, 46 :139–144, 2016.

[60] Htet Thazin Tike Thein and Khin Mo Mo Tun. An approach for breast cancer diagnosis classification using neural network. *Advanced Computing. An International Journal (ACIJ)*, 6 :1–11, 2015.

[61] Jian-sheng Guan, Lo-Yi Lin, Guo-li Ji, Chih-Min Lin, Tien-Loc Le, and Imre J. Rudas. Breast tumor computer-aided diagnosis using self-validating cerebellar model neural networks. *Acta Polytechnica Hungarica*, 13 :39–52, 2016.

[62] U. Karthik Kumar, M.B. Sai Nikhil, and K. Sumangali. Prediction of breast cancer using voting classifier technique. In *2017 IEEE international conference on smart technologies*

and management for computing, communication, controls, energy and materials (ICSTM), pages 108–114. IEEE, 2017.

[63] Dishant Mittal, Dev Gaurav, and Sanjiban Sekhar Roy. An effective hybridized classifier for breast cancer diagnosis. In *2015 IEEE international conference on advanced intelligent mechatronics (AIM)*, pages 1026–1031. IEEE, 2015.

[64] Emina Aličković and Abdulhamit Subasi. Breast cancer diagnosis using ga feature selection and rotation forest. *Neural Computing and applications*, 28 :753–763, 2017.

[65] Bichen Zheng, Sang Won Yoon, and Sarah S. Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41 :1476–1482, 2014.

[66] Rekh Ram Janghel, Anupam Shukla, Ritu Tiwari, and Rahul Kala. Breast cancer diagnosis using artificial neural network models. In *The 3rd International Conference on Information Sciences and Interaction Sciences*, pages 89–94. IEEE, 2010.

[67] Vikas Chaurasia and Saurabh Pal. Data mining techniques : to predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing IJCSMC*, 3 :10–22, 2014.

[68] Mehrbakhsh Nilashi, Othman Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. A knowledge-based system for breast cancer classification using fuzzy logic method. *Telematics and Informatics*, 34(4) :133–144, 2017.

[69] Cuong Nguyen, Yong Wang, and Ha Nam Nguyen. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. 2013.

[70] Jacob Dheeba, N Albert Singh, and S Tamil Selvi. Computer-aided detection of breast cancer on mammograms : A swarm intelligence optimized wavelet neural network approach. *Journal of biomedical informatics*, 49 :45–52, 2014.

[71] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv :1606.05718*, 2016.

[72] Sutong Wang, Yuyan Wang, Dujuan Wang, Yunqiang Yin, Yanzhang Wang, and Yaochu Jin. An improved random forest-based rule extraction method for breast cancer diagnosis. *Applied Soft Computing*, 86 :105941, 2020.

[73] Mesut Toğaçar, Burhan Ergen, and Zafer Cömert. Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders. *Medical hypotheses*, 135 :109503, 2020.

[74] Mesut Toğaçar, Kutsal Baran Özkurt, Burhan Ergen, and Zafer Cömert. Breastnet : a novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. *Physica A : Statistical Mechanics and its Applications*, 545 :123592, 2020.

[75] Moloud Abdar and Vladimir Makarenkov. Cwv-bann-svm ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement*, 146 :557–570, 2019.

[76] Abir Alharbi and F Tchier. Using a genetic-fuzzy algorithm as a computer aided diagnosis tool on saudi arabian breast cancer database. *Mathematical biosciences*, 286 :39–48, 2017.

[77] Reza Rasti, Mohammad Teshnehlab, and Son Lam Phung. Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. *Pattern Recognition*, 72 :381–390, 2017.

[78] Ümit Budak, Zafer Cömert, Zryan Najat Rashid, Abdulkadir Şengür, and Musa Çıbuk. Computer-aided diagnosis system combining fcn and bi-lstm model for efficient breast cancer detection from histopathological images. *Applied Soft Computing*, 85 :105765, 2019.

[79] Sami Ekici and Hushang Jawzal. Breast cancer diagnosis using thermography and convolutional neural networks. *Medical hypotheses*, 137 :109542, 2020.

[80] Ruholla Jafari-Marandi, Samaneh Davarzani, Maryam Soltanpour Gharibdousti, and Brian K Smith. An optimum ann-based breast cancer diagnosis : Bridging gaps between ann learning and decision-making goals. *Applied Soft Computing*, 72 :108–120, 2018.

[81] Shuo Liu, Jinshu Zeng, Huizhou Gong, Hongqin Yang, Jia Zhai, Yi Cao, Junxiu Liu, Yuling Luo, Yuhua Li, Liam Maguire, et al. Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach. *Computers in biology and medicine*, 92 :168–175, 2018.

[82] Na Liu, Er-Shi Qi, Man Xu, Bo Gao, and Gui-Qiu Liu. A novel intelligent classification model for breast cancer diagnosis. *Information Processing & Management*, 56(3) :609–623, 2019.

[83] Bibhuprasad Sahu, Sachi Mohanty, and Saroj Rout. A hybrid approach for breast cancer classification and diagnosis. *EAI Endorsed Transactions on Scalable Information Systems*, 6(20), 2019.

[84] Meerja Akhil Jabbar. Breast cancer data classification using ensemble machine learning. *Engineering and Applied Science Research*, 48(1) :65–72, 2021.

[85] Ankur Makwana and Jaymin Patel. Decision support system for heart disease prediction using data mining techniques. *International Journal of Computer Applications*, 117(22) :1–5, 2015.

[86] B Subanya and RR Rajalaxmi. Feature selection using artificial bee colony for cardiovascular disease classification. In *2014 International conference on electronics and communication systems (ICECS)*, pages 1–6. IEEE, 2014.

[87] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K Mago. Building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1377–1382. IEEE, 2016.

[88] Renu Narain, Sanjai Saxena, and Achal Kumar Goyal. Cardiovascular risk prediction : a comparative study of framingham and quantum neural network based approach. *Patient preference and adherence*, 10 :1259, 2016.

[89] B Venkatalakshmi and MV Shivsankar. Heart disease diagnosis using predictive data mining. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(3) :1873–7, 2014.

[90] Ahmed Hamed, Ahmed Sobhy, and Hamed Nassar. Accurate classification of covid-19 based on incomplete heterogeneous data using a knn variant algorithm. *Arabian Journal for Science and Engineering*, 46(9) :8261–8272, 2021.

[91] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121 :103792, 2020.

[92] Vinay Kumar Reddy Chimmula and Lei Zhang. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals*, 135 :109864, 2020.

[93] Jordi Laguarta, Ferran Hueto, and Brian Subirana. Covid-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, 1 :275–281, 2020.

[94] Gonçalo Marques, Deevyankar Agarwal, and Isabel de la Torre Díez. Automated medical diagnosis of covid-19 through efficientnet convolutional neural network. *Applied soft computing*, 96 :106691, 2020.

[95] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *European radiology*, 31(8) :6096–6104, 2021.

[96] RG Babukarthik, V Ananth Krishna Adiga, G Sambasivam, D Chandramohan, and J Amudhavel. Prediction of covid-19 using genetic deep learning convolutional neural network (gdcnn). *Ieee Access*, 8 :177647–177666, 2020.

[97] Moutaz Alazab, Albara Awajan, Abdelwadood Mesleh, Ajith Abraham, Vansh Jatana, and Salah Alhyari. Covid-19 prediction and detection using deep learning. *International Journal of Computer Information Systems and Industrial Management Applications*, 12(June) :168–181, 2020.

[98] Talha Burak Alakus and Ibrahim Turkoglu. Comparison of deep learning approaches to predict covid-19 infection. *Chaos, Solitons & Fractals*, 140 :110120, 2020.

[99] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Ruixuan Wang, Huiying Zhao, Yutian Chong, et al. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(6) :2775–2780, 2021.

[100] Fatima M Salman, Samy S Abu-Naser, Eman Alajrami, Bassem S Abu-Nasser, and Belal AM Alashqar. Covid-19 detection using artificial intelligence. 2020.

[101] Halgurd S Maghdid, Aras T Asaad, Kayhan Zrar Ghafoor, Ali Safaa Sadiq, Seyedali Mirjalili, and Muhammad Khurram Khan. Diagnosing covid-19 pneumonia from x-ray and ct images using deep learning and transfer learning algorithms. In *Multimodal image exploitation and learning 2021*, volume 11734, pages 99–110. SPIE, 2021.

[102] Farhan Mohammad Khan, Akshay Kumar, Harish Puppala, Gaurav Kumar, and Rajiv Gupta. Projecting the criticality of covid-19 transmission in india using gis and machine learning methods. *Journal of Safety Science and Resilience*, 2(2) :50–62, 2021.

[103] Jiangpeng Wu, Pengyi Zhang, Liting Zhang, Wenbo Meng, Junfeng Li, Chongxiang Tong, Yonghong Li, Jing Cai, Zengwei Yang, Jinhong Zhu, et al. Rapid and accurate identification of covid-19 infection through machine learning based on clinical available blood test results. *MedRxiv*, 2020.

[104] Mohammadreza Nemati, Jamal Ansary, and Nazafarin Nemati. Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns*, 1(5) :100074, 2020.

[105] Shaoping Hu, Yuan Gao, Zhangming Niu, Yinghui Jiang, Lao Li, Xianglu Xiao, Minhao Wang, Evandro Fei Fang, Wade Menpes-Smith, Jun Xia, et al. Weakly supervised deep learning for covid-19 infection detection and classification from ct images. *IEEE Access*, 8 :118869–118883, 2020.

[106] Ali Abbasian Ardakani, Alireza Rajabzadeh Kanafi, U Rajendra Acharya, Nazanin Khadem, and Afshin Mohammadi. Application of deep learning technique to manage covid-19

in routine clinical practice using ct images : Results of 10 convolutional neural networks. *Computers in biology and medicine*, 121 :103795, 2020.

[107] Mucahid Barstugan, Umut Ozkaya, and Saban Ozturk. Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint arXiv :2003.09424*, 2020.

[108] Md Zahangir Alom, MM Rahman, Mst Shamima Nasrin, Tarek M Taha, and Vijayan K Asari. Covid_mtnet : Covid-19 detection with multi-task deep learning approaches. *arXiv preprint arXiv :2004.03747*, 2020.

[109] Chuansheng Zheng, Xianbo Deng, Qiang Fu, Qiang Zhou, Jiapei Feng, Hui Ma, Wenyu Liu, and Xinggang Wang. Deep learning-based detection for covid-19 from chest ct using weak label. *MedRxiv*, 2020.

[110] Gergo Pinter, Imre Felde, Amir Mosavi, Pedram Ghamisi, and Richard Gloaguen. Covid-19 pandemic prediction for hungary ; a hybrid machine learning approach. *Mathematics*, 8(6) :890, 2020.

[111] LJ Muhammad, Ebrahem A Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty, and Ibrahim Alh Mohammed. Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. *SN computer science*, 2(1) :1–13, 2021.

[112] Harsh Panwar, PK Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, and Vaishnavi Singh. Application of deep learning for fast detection of covid-19 in x-rays using ncovnet. *Chaos, Solitons & Fractals*, 138 :109944, 2020.

[113] Shashank Vaid, Reza Kalantar, and Mohit Bhandari. Deep learning covid-19 detection bias : accuracy through artificial intelligence. *International Orthopaedics*, 44(8) :1539–1542, 2020.

[114] Syed Hammad Hussain Shah, Ole-Martin Steinnes, Eirik Gribbestad Gustafsson, and Ibrahim A Hameed. Multi-agent system based mobile help desk and monitoring of safety measures to combat covid-19 and future pandemics. In *2021 International Conference on Artificial Intelligence (ICAI)*, pages 80–85. IEEE, 2021.

[115] Giuseppe Antonio Nanna, Nicola Flavio Quatraro, and Berardina De Carolis. A multi-agent system for simulating the spread of a contagious disease. In *WOA*, volume 1613, page 119, 2020.

[116] Yaroslav Vyklyuk, Mykhailo Manylich, Miroslav Škoda, Milan M Radovanović, and Marko D Petrović. Modeling and analysis of different scenarios for the spread of covid-19 by using the modified multi-agent systems–evidence from the selected countries. *Results in Physics*, 20 :103662, 2021.

[117] Samir Kumar Bandyopadhyay and Shawni Dutta. Artificial intelligence based study on analyzing of habits and with history of diseases of patients for prediction of recurrence of disease due to covid-19. 2020.

[118] Ali Akbar Sadat Asl, MM Ershadi, Shahabeddin Sotudian, X Li, and S Dick. Fuzzy expert systems for prediction of icu admission in patients with covid-19. *Intelligent Decision Technologies*, (Preprint) :1–10, 2022.

[119] Lauren M Boden, Stephanie A Boden, Ajay Premkumar, Michael B Gottschalk, and Scott D Boden. Predicting likelihood of surgery before first visit in patients with back and lower extremity symptoms : A simple mathematical model based on more than 8,000 patients. *Spine*, 43(18) :1296–1305, 2018.

[120] Kjetil Søreide, Kenneth Thorsen, and Jon Arne Søreide. Predicting outcomes in patients with perforated gastroduodenal ulcers : artificial neural network modelling indicates a highly complex disease. *European Journal of Trauma and Emergency Surgery*, 41(1) :91–98, 2015.

[121] Nyssa Hunt, Andrew Carroll, Thomas P Wilson, et al. Spatiotemporal analysis and predictive modeling of rabies in tennessee. *Journal of Geographic Information System*, 10(01) :89, 2018.

[122] C Sharmila Devi, G Geetha Ramani, and J Arun Pandian. Intelligent e-healthcare management system in medicinal science. *International Journal of PharmTech Research*, 6(6) :1838–1845, 2014.

[123] Kaberi Das, Debahuti Mishra, and Kailash Shaw. A metaheuristic optimization framework for informative gene selection. *Informatics in Medicine Unlocked*, 4 :10–20, 2016.

[124] Golnaz Sahebi, Amin Majd, Masoumeh Ebrahimi, Juha Plosila, and Hannu Tenhunen. A reliable weighted feature selection for auto medical diagnosis. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, pages 985–991. IEEE, 2017.

[125] Talayeh Razzaghi, Ilya Safro, Joseph Ewing, Ehsan Sadrfaridpour, and John D Scott. Predictive models for bariatric surgery risks with imbalanced medical datasets. *Annals of Operations Research*, 280(1) :1–18, 2019.

[126] S Shanthi and N Rajkumar. Lung cancer prediction using stochastic diffusion search (sds) based feature selection and machine learning methods. *Neural Processing Letters*, 53(4) :2617–2630, 2021.

[127] Tuncay Bayrak and Hasan Ogul. Computer-aided diagnosis of sleep apnea using gene expression. *Health and Technology*, 11(4) :941–952, 2021.

[128] U.s. cancer statistics working group. united states cancer statistics : 19992008 incidence and mortality web-based report. atlanta (ga) :2012. department of health and human services, centers for disease control and prevention, and national cancer institute.

[129] Djihane Houfani, Sihem Slatnia, Okba Kazar, Noureddine Zerhouni, Abdelhak Merizig, and Hamza Saouli. Machine learning techniques for breast cancer diagnosis : literature review. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 247–254. Springer, 2020.

[130] Djihane Houfani, Sihem Slatnia, Okba Kazar, Hamza Saouli, and Abdelhak Merizig. Artificial intelligence in healthcare : a review on predicting clinical needs. *International Journal of Healthcare Management*, 15(3) :267–275, 2022.

[131] Sachin S Kamble, Angappa Gunasekaran, Milind Goswami, and Jaswant Manda. A systematic perspective on the applications of big data analytics in healthcare management. *International Journal of Healthcare Management*, 2018.

[132] M Ciotti, M Ciccozzi, A Terrinoni, WC Jiang, CB Wang, and S Bernardini. La pandemia de covid-19. *Crit Rev Clin Lab Sci*, 17 :365–88, 2020.

[133] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+\Wisconsin+(Diagnostic). [En ligne ; Consulté le 20 septembre 2022].

[134] Djihane HOUFANI, Sihem SLATNIA, Okba KAZAR, Noureddine ZERHOUNI, Hamza SAOULI, and Ikram REMADNA. Breast cancer classification using machine learning techniques : a comparative study. *Medical Technologies Journal*, 4(2) :535–544, 2020.

[135] Djihane Houfani, Sihem Slatnia, Okba Kazar, Ikram Remadna, Hamza Saouli, Guadalupe Ortiz, and Abdelhak Merizig. An improved model for breast cancer diagnosis by combining pca and logistic regression techniques. *International Journal of Healthcare Management, Elsevier*, 2023.

[136] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). [En ligne ; Consulté le 20 septembre 2022].

[137] V Nanda Gopal, Fadi Al-Turjman, R Kumar, L Anand, and M Rajesh. Feature selection and classification in breast cancer prediction using iot and machine learning. *Measurement*, 178 :109442, 2021.

[138] Kemal Polat and Salih Güneş. Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, 17(4) :694–701, 2007.

[139] https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset. [En ligne ; Consulté le 20 juin 2022].

[140] Katia Sycara. Multi-agent systems. *Al Magazine*, 19 :79–92, 1998.

[141] Sercan Ö Arik and Tomas Pfister. Tabnet : Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

[142] Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9) :1315–1316, 2010.